University of Glamorgan

Prifysgol Morgannwg

The Use of Artificial Intelligence Techniques to Assist in the Valuation of Residential Properties

Owen Michael Lewis

School of Accounting and Mathematics

and School of the Built Environment

University of Glamorgan

Pontypridd

A thesis submitted in partial fulfilment of the requirements of the University of Glamorgan for the degree of Doctor of Philosophy.

July 1999

Table of Contents

Table of Figures

Acknowledgements

I would like to express sincere thanks to my supervisors at the University of Glamorgan, Dr. J.A.Ware, Professor.S.A.Gronow and Mr. D.H. Jenkins, and to Mr.H.James formerly of Portsmouth University, for the support, encouragement, guidance and constructive criticism they have provided throughout this research project.

Thanks also to my fellow research students for their camaraderie and support throughout my time at the University of Glamorgan.

Finally, thanks to my wife Georgina and children Jessica and Daniel for their love, support and patience.

Certificate of Research

This is to certify that, except where specific reference is made, the work presented in this thesis is the result of the investigation undertaken by the candidate.

Candidate .....

Director of Studies .....

Declarations

This is to certify that neither this thesis or any part of it has been presented or is being currently submitted in candidature for any other degree other than the degree of Doctor of Philosophy of the University of Glamorgan.

Candidate .....

Abstract

The Use of Artificial Intelligence Techniques to Assist in the Valuation of Residential Properties

Owen Michael Lewis

This thesis documents the research that has led to the development of a number of methodologies for combining existing artificial intelligence and statistical techniques into a form appropriate for the development of an intelligent appraisal system for use in the residential property appraisal profession. The methodologies illustrate how regression based appraisal models, previously restricted to homogeneous data, can be applied to heterogeneous data without significant loss in accuracy. The majority of research, previous to this, has addressed this problem by manually selecting homogeneous sub-regions from a heterogeneous parent region. However, the main drawback with this approach is that the segregation of parent regions into sub-regions relies upon a significant amount of a priori knowledge pertaining to the location of the property being valued. The requirement for a commercial residential property appraisal system is one that given sufficient training evidence can automatically learn how to value a property in any region and be able to modify this knowledge over time.

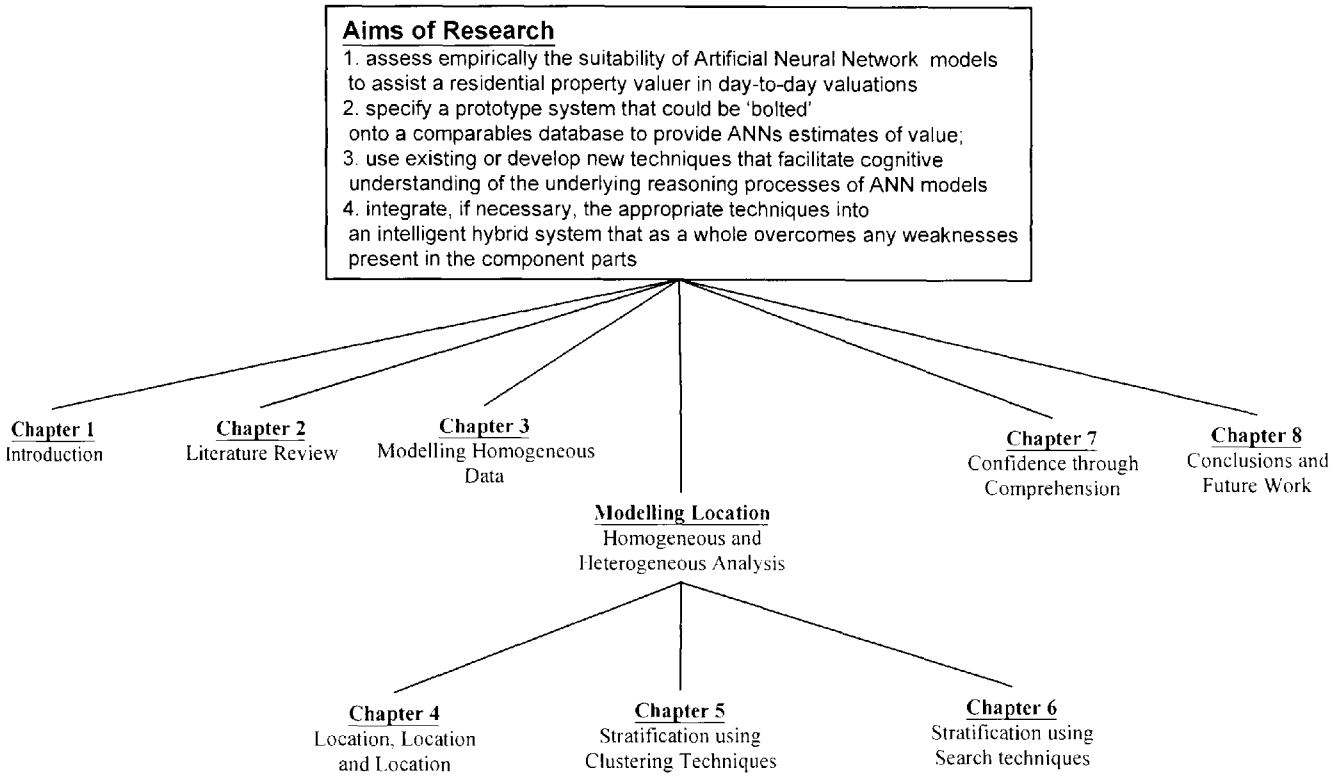
Two methodologies are proposed within the thesis to address this requirement. The first, using a technique known as the Kohonen Self Organising Map, makes an assumption that residential properties that share sufficient characteristics can be appraised using the same function. The Kohonen Self Organising Map is used to cluster properties with respect to their property characteristics and locational characteristics represented using a mortgage transaction database and UK Census statistics. Aptness of each cluster to define a homogeneous subset suitable to train a regression model, such as multiple regression analysis or a neural network, is estimated using a form of 'nearest neighbour' analysis. The second methodology, improves on the previous by transforming the static 'cluster then observe' solution to a more dynamic one using a Genetic Algorithm to evolve good clusters from those that at first inspection were mediocre.

Another issue that has hindered the development of intelligent residential property appraisal systems has been the inability of such models to express their underlying functional form. This is addressed from two perspectives in this thesis: Rules are derived that describe the characteristic make-up of the formed clusters and, alternative modelling techniques are used to generate the final training models that are able to express their functional form as a set of induced rules.

The work contained within this thesis demonstrates the feasibility of such an automatic stratification approach. Also, the research illustrates that by observing the characteristics of the generated clusters formed, a useful insight into both the underlying reasoning of the generated models and also of the locational and financial makeup of the subject location can be gained.

Thesis Outline

Chapter 1 provides the background and rational for the research. A critique of appraisal techniques published in academia and present in practice is given in Chapter 2. Following, a programme of empirical investigation is detailed for both homogeneous (Chapter 3) and heterogeneous (Chapter 4) market analysis. Methods considered for inclusion in an automated appraisal model are detailed in Chapters 5 and 6 with Chapter 7 providing techniques for inspiring confidence through model comprehension. The thesis concludes in Chapter 8 with suggestions made for further work.



1. INTRODUCTION

This chapter sets the scene for the remainder of the thesis and provides a rationale for the research before giving an overview of the aims and objectives of the work undertaken.

1.1 Background

The value of a residential property reflects its relative position in the demand and supply equilibrium, maximising the return to the vendor and at the same time satisfying the expectations of the buyer. Given the capital requirements, an intermediary is often used to allow the buyer to meet the contract agreed with the vendor. As a return on this investment, the intermediary demands a with-interest repayment spread over an agreed period. In providing this service the intermediary or lender, which is normally a bank or building society, accepts a certain amount of risk and must assess the level of risk before entering into such an agreement. To assess this risk the lender must establish the true open market value (OMV) of the property and compare this with the amount of capital the buyer wishes to borrow. The buyer also accepts a certain amount of risk. The risk in this case is that the value of his property may fall below the commencement value of the contract - a situation known as 'negative equity'. The value used as a benchmark to assess the level of risk in both cases is set by professional valuers (in many cases working for the lending institution or connected in some way - for example, through a subsidiary company).

1.2 Residential Property Appraisal

The conventional method for valuing a residential property is the method of Direct Capital Comparison (DCC) (See Mackmin, 1994). DCC involves selecting properties comparable to the subject property sold in an 'open market'. The valuer makes an *"allowance in money terms"* (Millington, 1994) for any differences between the subject property and the comparable properties. Valuers rely heavily upon experience, which

"produces an empathy for movements in the market, and allows the experienced valuer to reconcile differences among comparable sales evidence and so produce an accurate opinion of value". (Adair and McGreal, 1986)

1.3 Effects of Recent Appraisal Procedure

In the late 1980's, residential property prices in the UK rose to record heights. The property market, led by a frenzy of activity in London and the South-East, was in a boom period. Banks and building societies were constantly reaching new lending targets. However, as with most booms, there followed a bust. In 1989 the market collapsed and thousands of homeowners faced varying degrees of negative equity. It became clear to some professional appraisers, academics and homeowners, that the valuation methods and procedures used by lenders had failed to assess adequately the risk involved in what to most people is the largest capital investment of their lives.

In the aftermath of this deflated market, scrutiny has been placed on the way in which valuers and lending institutions arrive at their OMVs. Researchers in the UK, in the United States of America and elsewhere, have focused on the theory and application of current appraisal methods and considered alternative techniques to assist residential property valuers to converge to the true OMV.

1.4 Alternative Appraisal Techniques

In an attempt to furnish professional appraisers with consultation tools for use during the appraisal procedure, a number of alternative techniques are being considered. The most significant of which are:

- Multiple Regression Analysis (MRA)
- Expert Systems and Databases
- Linear Programming
- Artificial Neural Networks (ANNs)

These techniques have been applied with varying degrees of success by many researchers using residential property data from a number of countries. Although there have been many publications describing alternative valuation methods (see review in Chapter 2), much research and development work is still required to establish which method to use, how to apply it, and indeed whether the solution

should in fact be a hybrid. When this stage is reached, there is the further challenge of inspiring confidence amongst professional appraisers and lending institutions. The following sections provide an introduction to these techniques and their application in residential property appraisal.

Multiple Regression Analysis: Simple linear regression is concerned with the relationship between two variables X and Y, where the value of Y (the dependent variable) is dependent on the corresponding value of X (the independent variable). Multiple regression analysis is used to describe the relationship between one dependent variable Y and many independent variables $X_1..X_n$. For a system of the form:

$$Y = a + bX_1 + cX_2 + dX_3 + ... + zX_n, + Error \quad (1.1)$$

MRA finds the best numerical values of {a,b,c,..,z} so as to minimise the squared difference between the actual values A and the predicted values P. This can be achieved a number of ways, with the most popular being:

- Standard MRA;
- Forward Stepwise Regression (variables entered one-by-one into the model if statistically significant);
- Backward Stepwise Regression (initially all variables entered into model, those that are not statistically significant are removed);
- Others (constrained, ridge, non-linear, constrained non-linear)

For modelling residential property appraisal, the dependent variable is usually set at the property value (or sometimes sale price) and the independent variables are those attributes that describe a property and its locational features.

Expert Systems (ES) and Expert Databases (ED): ES are computer-based techniques that emulate the procedures taken by an expert in a well-defined domain. An ES has three basic components:

Fact Base: This contains all the facts upon which an expert can make his judgement.
Knowledge Base: This contains all the procedures and rules that an expert can apply to the given facts in order to reach a decision.

Inference Engine: Here appropriate rules from the knowledge base are selected and applied given the current state of the facts base, often resulting in new facts being determined.

ES are able to replicate the actions of an expert, provided the expert can articulate his expertise. This approach can be used to model the comparable adjustment process using a database system, resulting in an 'expert database' system.

Linear Programming (LP): LP is an optimisation technique. That is, a technique which given a set of parameters and constraints, will maximise or minimise an objective function. An example of an objective function is:

ValueOfHouse = A*ValueOfBedroom + B*ValueOfDryRot + .. + Z*ValueOfGarage

Here, the LP approach could be used to maximise ValueOfHouse by maximising and minimising the independent variables bound within the constraints (upper and lower bound) of comparable properties.

Artificial Neural Networks: ANNs, like MRA, model the relationship between the independent variables describing the subject property and its dependent value. The difference between the two techniques is in the structure of the model. ANNs are a distributed array of highly interconnected processing elements. Each connection has an associated strength or weight. Processing at each node involves a weighted sum of each connection forming a single input to a non-linear transfer function. The output from each node becomes the input to successive nodes.

1.5 Research Aim

The aim of this research was to assess the potential for using appropriate Artificial Intelligence (AI) techniques as tools to assist in the appraisal of residential properties. In particular the research sought to:

1. assess empirically the suitability of Artificial Neural Network models to assist a residential property valuer in day-to-day valuations;
2. specify a prototype system that could be 'bolted' onto a comparables database to provide ANNs estimates of value;

3. use existing (or develop new) techniques that facilitate cognitive understanding of the underlying reasoning processes of ANN models;
4. if necessary, integrate, the appropriate techniques into an intelligent hybrid system that as a whole overcomes any weaknesses present in the component parts;

The rationale for this research came from an established research record at the forefront of residential property appraisal in both academic and professional circles. Inspired by recent computing developments the research team at the University of Glamorgan began to broaden its perspective and consider the use of AI techniques as valuation instruments. Attention was originally paid to expert systems (Gronow and Scott, 1987) and later to expert databases (Jenkins, 1992). A number of leading papers highlighted the potential for automating the data collection process during mortgage surveys (Jenkins, 1992) and formalising the DCC method of valuation (Gronow, et al, 1996).

The research base then broadened to include techniques for estimating the underlying functions present in mortgage approval data. Drawing on the foundations set by significant contributions in both Multiple Regression Analysis (Antwi, 1995; Adair and McGreal, 1987) and Artificial Neural Networks (Evans, et al, 1992; Do and Grudnitski, 1992; Tay and Ho, 1992; Borst, 1991; Borst, 1993), and, using funding from a Realising Our Potential Award under the auspices of the ESRC, the research aims presented in this thesis were born.

1.6 Structure of Thesis

A critique of appraisal techniques published in academia and present in practice is given in Chapter 2. Chapter 3 and Chapter 4 present homogeneous and heterogeneous market analysis, respectively. Methods considered for inclusion in an automated appraisal model are detailed in Chapters 5 and 6 with Chapter 7 providing techniques for inspiring confidence through model comprehension. The thesis concludes in Chapter 8 where suggestions are made for further work.

2. LITERATURE REVIEW

This chapter contains a review of relevant literature pertaining to current and alternative techniques for appraising residential properties. Each technique is introduced and its current role in practice and academia is stated, followed by a summary of its relative merits and weaknesses as discussed in professional and academic literature

2.1 Introduction

Within the United Kingdom valuation of residential properties for mortgage purposes is undertaken using the method of Direct Capital Comparison (DCC), alternatively called the comparative or market sales approach. Such valuations could however be performed using the income or investment approach; contractors or cost approach; or, the residual or developers approach (Mackmin, 1994). Despite the widespread use of the DCC method, it has come under considerable pressure and criticism following the collapse of the housing market in the late 1980's. Both the logic and the application being scrutinised, with academics reporting that the method is imprecise, ambiguous (Wiltshaw, 1991a) and weak (Jenkins, 1992). Previous research has focused primarily on the formalisation of the DCC method and the development of alternative valuation techniques.

2.2 Direct Capital Comparison

This method is the most frequently used in residential property appraisal as it is deemed to provide a true market valuation. There are four steps involved in its application:

- Select comparables;
- Extract, confirm and analyse comparable sale prices;
- Adjust sale prices for notable differences;
- Formulate an opinion of Open Market Value (OMV) for the subject property.

(Mackmin, 1994)

In practice valuers rely upon data banks of knowledge relating to past transactions processed within the valuation establishment. Adjustments are normally made for differences in sale dates, location, condition and accommodation.

Evidence suggests (Almond, et al, 1997) that leading residential surveying firms are beginning to make use of database technology which speeds up the process of sorting through records of past transactions in order to find suitable comparables. However, even with the recent introduction of IT to the profession, good comparables are sparse - especially for unusual properties. This problem is compounded in smaller practices in England and Wales, as information on sales held by the Land Registry is not readily available. Wiltshaw (1993) acknowledges this problem, reporting that the confidentiality of property prices in England and Wales leads to a *"thin market in traded properties"* that often yield at best only a *"few potential comparables"*. This has led valuers to compare location and environmental benefits as well as physical attributes for more unusual properties (Mackmin, 1994).

Critical studies of the DCC method have revealed a number of shortfalls in both application and theory. The size of samples used by valuers has been questioned. Adair and McGreal (1987) suggest an experienced valuer often relies on three or less comparables, a sample size that Wiltshaw (1991b) states is *"too small for an econometric valuation"*. Although, from a practitioner's point of view, having one good comparable is far better than half a dozen sales that require significant adjustment (Mackmin, 1994). More worrying than sample size, however, is the influence of transaction price on the value returned by the valuer. In a study of over 100,000 properties, 65% of observations showed an exact match and 90% of values returned by the valuer fell within 5% of the transaction price (Gronow, et al, 1996). Clearly revealing the transaction price prior to the mortgage survey tends to bias the value returned by the valuer.

Further problems lie in the application of the DCC method, with many valuers dispensing with a formal comparable grid in favour of carrying out ad hoc adjustments. Here there is a danger of valuers relying on comparables that are unrepresentative and outdated.

Clearly, the literature highlights a number of weaknesses in the DCC method, with these weaknesses remaining relatively unchallenged from within the valuation profession. There is also a notable deficiency in attention to this method in under-graduate texts (Almond, et al, 1997). Given the poor application and lack of statistical inference, research has begun to address the limiting factors of the method and also to pioneer complementary valuation techniques for the valuer to use as additional evidence (See Almond, 1999).

2.3 Multiple Regression Analysis

Drawing from statistical analysis, the most structured approach for inducing a model from process data is regression. The principle being that coefficients are optimised - by reducing a global error term on the process data - and then used to formulate an aggregate value, based on the sum of the individual component values. This approach, generally termed Hedonic Pricing, has been applied to residential property data to provide the valuer with a statistically derived value to complement the existing DCC value. Furthermore, the coefficients can give an explicit indication of the influence each attribute has on the final property value (Miller, 1982).

Hedonic pricing studies based primarily on linear multiple regression analysis are numerous in the UK and the US. However, despite many introductory texts (Adair and McGreal, 1987; Antwi, 1995), UK interest has remained within academia, although, the method is used widely for compiling house price indices (Flemming and Nellis, 1994).

With modern statistical packages such as SAS or SPSS, a simple MRA model can be created in minutes. The model can be used to estimate the value of a new case or the coefficients analysed to gain an understanding of their net influence. For example, Colwell and Foley (1979) considered the effect of electricity transmission lines on value, Guntermann and Colwell (1983) the proximity of schools, and Pennington, et al (1990) the impact of aircraft noise.

However, despite its simplicity and statistical harnessing, the elementary approach has severe restrictions. As with most statistical tools, the sample size must be significant and hence the availability of good data is important. Shenkel (1978) echoed by Adair and McGreal (1987) states that 100 comparable sales giving good

information on size, location, physical and neighbourhood characteristics is a prerequisite for the use of MRA. This of course is a problem common to most data-orientated modelling techniques. Perhaps more significant is the underlying structure of MRA requiring the relationships amongst property attributes to be additive and linear. Adair, et al, (1996) suggests this functional form has a *"wide application in real estate"*. However, Bruce and Sundell (1977), stating that, *"the issues involved in property valuation are much too complex for the simple additive theory on which it [MRA] is based"*, challenge this. Violation of the assumptions of linearity render basic significance tests and linear regression analysis unusable. Of course, non-linear multiple regression analysis is available in most statistical analysis packages. The pre-determination of the underlying functional form (e.g. polynomial) renders this method extremely difficult especially for multidimensional data (see example in Goulden, 1989).

Furthermore, the selection of suitable attributes to form the regression equation can cause significant problems. MRA is affected by multicollinearity, a situation where two or more variables are highly correlated. Lessinger (1969) suggests that multicollinearity *is "more often the rule rather than the exception"*, and Newell (1982) reports that multicollinearity is a *"common feature in property data"*. A typical example of this phenomenon is the number of bedrooms increasing with lot size (Li and Brown, 1980). MRA produces inaccurate predictions using multicollinear data and regression coefficients have *"no practical interpretation"* (Newell, 1982).

There are methods of overcoming multicollinearity, the most obvious being the removal of one of the correlated pair or transforming the variables into ratios. A more sophisticated method of reducing the correlated variables without losing representativeness is to extract orthogonal linear components, otherwise known as principle components. Bourassa, et al, (1997), considered this approach in their study of housing markets in Sydney and Melbourne, Australia. Using principle component analysis, the authors were safe to assume normality and hence use MRA with confidence. However, as the attributes no longer relate to real property characteristics, the subsequent interpretation of the coefficients becomes very subjective, with the authors admitting that some factors were *"difficult to characterise"* (Bourassa, et al, 1997).

Alternatively, MRA can be substituted by Ridge Regression. This technique, although similar to MRA, attempts to overcome multicollinearity by adding a bias estimate for each regression coefficient. Newell (1982) reports a general improvement over MRA, but notes that the method requires expert judgement in manually setting the bias to ensure *"stable property characteristics"* (Newell, 1982).

In addition to multicollinearity, MRA analysis is further weakened with skewness distributions. Although Adair and McGreal (1987) point out that mathematical transformations can force skewed data to satisfy the assumptions of the model, they go on to suggest that extreme values should be excluded.

Despite these weaknesses, MRA has formed the major contribution to alternative valuation literature in the UK and US, with its simplicity, explicit structure and statistical basis appealing to many researchers. Dodgson (1989) feels that MRA provides an *"extremely promising and cost-effective"* way of valuing residential properties. With Ashton (1972) suggesting that MRA is an appraisal technique that *"will inevitably become, if it already has not done so, an invaluable complement to the appraiser's existing tool kit"* (Ashton, 1972). However, Adair and McGreal (1987) warn that MRA has *"severe restrictions"* if applied to heterogeneous data unless good indicators of location can be generated.

2.4 Linear Programming

Wiltshaw (1991a) states that there is a need for a *"reconstruction [of valuation methodology] which provides a sound algebraic and arithmetic framework"*. Further, any new method should not make any *"additional data demands"* (Wiltshaw, 1991a).

By redefining the comparable method as a set of non-homogeneous equations, Wiltshaw (1991a) theorises that a solution for these simultaneous equations can be achieved by finding the maximum value in the solution space using the Simplex method. Thus placing the valuation *"firmly in the realm of mathematics"* (Wiltshaw, 1991a). However the suggested method (linear programming LP) - in common with MRA - assumes a linear relation between the property characteristics and the dependent variable and is therefore subject to the same limitations.

In a subsequent commentary paper (Matysiak, 1991), the suitability of the LP technique as a valuation instrument is questioned. A warning is issued concerning a potential compounding problem where LP based valuations are fed into new LP models forcing a *"linear dependence between values"* Matysiak (1991).

More importantly, Matysiak points out that in order to describe a feasible region for the Simplex method to investigate, the number of equations (comparables) must be greater than the number of variables (property attributes). The following scenario is postulated:

"...suppose that we define five variables for five comparables and use the Simplex algorithm to determine the implied factor prices P1 to P5, and then use these prices to value the subject property. It is subsequently shown that the 'maximum' price arrived at in this manner was way out and litigation ensues. In his defence, the valuer argues that the approach was sound 'But your honour, we used a deterministic non-homogeneous set of linear equations.' The judge replies, 'Yes, but as a professional valuer, do you not feel that you should, at least, have considered the number of bedrooms as relevant to your valuation?' 'Yes', the valuer replies, 'the number of bedrooms was important, but unfortunately this would have made the number of variables greater than the number of equations, so I had to drop the bedrooms.' The judge rules, 'I find against you.'" Matysiak, (1991).

A simple examination of the Simplex algorithm, which attempts to select an optimal vertex formed by the constraining equations (comparables), supports Matysiak's statement. The Simplex method reduces a set of n equations in m unknowns where m > n to a set of n equations in n unknowns, thus excluding some of the original m unknowns. This results in a situation where potentially important variables are omitted from the final model.

In a reply to Matysiak's (1991) commentary paper, Wiltshaw (1991b) discounts these criticisms, saying they are *"irrelevant"* and *"flawed"*. In defence, Witlshaw (1991b) argues *that "there is no guarantee that a particular property characteristic will prove to be significant as an explanatory variable"* and often an alternative measure can be used. The discussion is continued by Matysiak (1992), who concludes *"I reiterate the point regarding LP as a mechanism, which lacks insight"* and drawing from an earlier

comment he says that LP *"can make no contribution towards understanding and enhancing the quality of the valuation process"*.

Undeterred, Wiltshaw (1993) restructured his argument and added probabilities into the LP model in order to estimate imperfect price information such as unknown realised sale prices of the selected comparables.

Despite the open and very critical nature of the discussion, clearly both Wiltshaw and Matysiak agree on the fact the traditional DCC method is flawed. Both supporting an *"econometric based approach to valuation"* (Wiltshaw, 1993).

2.5 Artificial Neural Networks

Besides the traditional statistical and algebraic techniques, another modelling technique emerged, this time from the field of Artificial Intelligence (AI). The technique's generic name being Artificial Neural Networks (ANNs) originating from studies of neural activities in the human brain. Although a diverse and complex field, ANNs have been used very selectively in real-estate appraisal, research focusing particularly on Multi-Layered Perceptron (MLP) networks trained using an Error Back-Propagation algorithm (See Chapter 3 for explanation).

Tazelaar (1989) describes ANNs as *"humanity's attempt to mimic the way the brain does things in order to harness its versatility and its ability to infer and intuit from incomplete or confusing information"*. ANNs are able to generalise from examples and have the ability to interpolate from previous learning. ANNs are often found working as pattern classifiers in areas *where "problem solutions are complex and difficult to specify, but which have an abundance of data from which a response can be learnt"* (DTI Guidelines, 1990). ANNs do not require an array of a priori knowledge, which in many cases is a prerequisite for MRA (Tay and Ho, 1992). ANNs learn by *"inducing the latent rules inherent in the training set of input and output patterns"* (Tay and Ho, 1992).

Two classes of training methods are used to determine the final trained ANN model: supervised and unsupervised. The principle difference being that supervised training requires the target solution for each training example to be known a priori, whereas unsupervised training does not. Thus an MLP network, which uses a supervised

training algorithm such as Error Back-Propagation, must have a training set including input features (e.g. property attributes) and output targets (e.g. property values).

ANNs are currently enjoying a renaissance, with research projects and commercial ventures growing daily. This popularity can be partly attributed to resurgence of Kolmogorov's Existence Theorem of 1957 provided by fairly recent advances in learning algorithms. Kolmogorov's Existence Theorem states:

"Given any continuous function f:[0,1]n Pm, f(x) = y, f can be implemented exactly by a three-layer feed-forward neural network having n fan-out processing elements in the first (x - input) layer, (2n+1) processing elements in the middle layer, and m processing elements in the top (y - output) layer." (Hecht-Nielsen, 1990)

This theorem combined with the 'BP (Back Propagation) Approximation Theorem' (Hecht-Nielsen, 1990 based on work by Rumelhart and McClelland, 1986 and Werbos, 1974) provides the *"scientific community with a confidence that an appropriate BP architecture for their specific problems must exist"* (Tay and Ho, 1992).

Universal approximation can be claimed by neural networks because of their ability to represent non-linear relationships. This is a major advantage over more conventional techniques such as MRA and LP, as ANN models are more general than a linear model and have the potential to be at least as general as a non-linear model. This is facilitated by the internal structure of the networks, with each neuron (or node) containing a mapping function that is often non-linear (for example, the sigmoid function).

This makes ANN technology particularly suitable to financial analysis where non-linear relationships are clearly evident amongst chaotic activity. Example applications of neural networks in the financial sector are: tactical asset allocation (Refenes, 1994); currency exchange rate prediction (Weigend, et al, 1991); and stock price prediction (Schöneburg, 1990). New articles appear with some regularity in related journals and professional magazines such as Technical Analysis of Stocks and Commodities and the futures traders' magazine Futures. In addition to this,

many books are available dealing with financial prediction and time series analysis using neural networks (e.g. Zirilli, 1997; Vemuri and Rogers, 1994).

Some evidence (Borst, 1991; Lawrence, 1992; Lam, 1996, Worzala, et al, 1995) to support the existence of non-linear relationships within property data exists. Furthermore, the concept of non-linear relationships in property data is intuitive. For example, the addition of a further bathroom to a property will probably increase its value, but the extent to which this increase is linear is limited. A terraced property with 8 bathrooms may even be worth less than a similar property with 2 bathrooms.

Neural networks are a relatively recent arrival to research in property appraisal, appearing first in the early 1990s. Borst, (1991) tends to be considered as the first study of neural networks for property appraisal, Borst himself stating that this approach is completely different from all previous research (Borst, 1991) (although, Tay and Ho (1991) published their paper on mass appraisal of residential apartments at about the same time). Estimated feedback models, which bare similarities with neural networks can, however, be traced back in the property field to the late 1970's (see Carbone and Longini, 1977; Sauter, 1985).

The early ANN studies, with the exception of Tay and Ho (1991,1992), were based exclusively in the USA (Danny P.H. Tay and David K.H. Ho are based in the National University of Singapore). Borst (1991), introduced concepts such as 'Artificial Neurons', 'Learning Methods' and 'Training Data Sets' to the professional valuation community. His informative introductory paper describes methods for turning comparables databases into training sets for the layered network structures. Non-linear aspects of property data are discussed and the various mapping functions available are also shown. The results presented lead Borst to conclude: *"neural networks deserve strong consideration by the assessment community"* (Borst, 1991).

Research by Lu and Lu (1992) highlights some of the advantages and disadvantages of using neuro-computing in valuation. In their favour, neural networks are able to: outperform traditional methods; account for complex interactions without the need for several years of experience required by their human counterparts; require no a priori knowledge or pre-programming; bypass the knowledge acquisition stage required to build a rule-based model and easily accommodate addition of new inputs such as

mortgage rates and construction cost indices. To their detriment, however, neural networks: are akin to a black-box[1] and reveal little of their processing logic; can take a very long time to train; require some expert knowledge to define the network structure and the feature representation (a summary of Lu and Lu, 1992).

Do and Grudnitski (1992) offer a comparison between an ANN model and a standard MRA model for real estate appraisal. They perform a non-biased comparison of both techniques using 105 independent test properties. Their results clearly indicate that the neural network model outperforms the MRA model. This analysis is further supported by a DCC model providing valuations that tend towards the neural network estimates as opposed to the respective MRA estimates (Do and Grudnitski, 1992). These results leave the authors feeling *"optimistic about the promise of neural networks ... for the appraisal of single-family dwellings"* (Do and Grudnitski, 1992) and suggest a similar approach could be taken to value commercial properties.

Tay and Ho (1991) also took this approach whilst modelling a Singapore residential property sub-market. This work is further developed in a second study which, in addition to reporting a complete neural network analysis of apartment values in Singapore, extends the analysis to consider the importance of the underlying structure of the trained neural network in providing transparency to an otherwise black-box model. Using a method developed by Garson (1991) the individual weight components of each artificial neuron were examined and used to rank the importance of each input feature. Using this technique, Tay and Ho (1991) concluded that floor area was the most influential factor affecting sale price.

In addition to the academic research, neural networks began to infiltrate professional practice in the USA. A company called HNC devised a neural network based appraisal system that successfully completed over 100,000 valuations in California and Florida. Each appraisal costing on average $35 - which is a saving of approximately 88% on previous methods (Schwartz, 1995). This technology has been implemented in the LoanProspector project run by the Federal Home Loan Mortgage Corp. made possible by a change in USA law allowing mass appraisal

[1] The phrase black-box is often used to describe neural networks as they are often perceived has having definite inputs and outputs but lack any underlying functional transparency.

techniques to be used to value properties of less than $250,000. Schwartz (1995) perceives that emerging technologies such as neural networks, fuzzy logic and the Internet could lead to a situation where the *"human approach to property valuation may be as outmoded as the buggy-whip"*.

The Journal of Property Valuation & Investment published the first application of neural networks to residential property appraisal in the UK. Evans, et al, (1992), encouraged by the work of Tay and Ho (1992), set about constructing an appraisal model for 34 post 1960 houses, selected from 14 streets in the Midlands. While the data set is notably small, the authors postulate that the data represent *"a very full coverage for England and Wales"* (Evans, et al, 1992) emphasising that private-sector valuers would typically have to base valuations on similar quantities of data. Care is taken in their paper to introduce UK based valuers to fundamental concepts involved in neural networks. Textual descriptors of each property are recoded into a numeric format required by an MLP network using a priori knowledge. This stage is very important as false rankings set during recoding can reduce the susceptibility of the data to effectively act as a training set. Borst (1991) advises that a single neuron should not be used to represent *"vastly different situations"* (Borst, 1991). A decision was made to discard certain features from the training set, for example: *"house number within any street was discarded as it was considered unlikely to be significant"* (Evans, et al, 1992).

The terminology used in Evans, et al, (1992) was very 'accessible' to valuation professionals and with the results showing prediction errors of between 5 and 7%. The conclusions are diplomatic, suggesting that such models could be used as *"an additional tool to speed normal valuations"*, allowing typical valuations to be performed by *"clerical staff, with a considerable saving in professional time"* and most appealing to mass appraisal requirements such as council tax valuations (Evans, et al, 1992).

By their own admission, however, the authors accept that research into residential property appraisal using neural networks is in its nascent stage with important challenges being the inclusion of *"realistic locational co-ordinates and transaction dates"* (Evans, et al, 1992) in order to allow the valuer to take better account of *"neighbourhood and inter-temporal effects"* (Evans, et al, 1992).

A study of real estate values in America attempts to progress the analysis to include locational characteristics (Borst, 1994). Data from a Geographical Information System (GIS) are used to adjust estimated values produced by a neural network. Error patterns (ratio of predicted value to actual value) are fed into a GIS system which allows the development of *"neighbourhood value correction factors"* (Borst, 1994).

Dodgson and Topham (1990) use postcodes to classify properties using CACI's neighbourhood classification scheme ACORN based on the 1991 Census. While Munro, (1986), consider the correlation between variations in house prices with variations in housing, neighbourhood and local authority policies on house improvement.

The inclusion of locational data in an ANN model may be sufficient to enhance the model's performance. However if the human model is considered, where valuers become experts in a particular location, perhaps a more appropriate solution would be to build 'expert networks' for each 'sub-market'. This approach was suggested by James (1994) - *"Another method of improving training is to determine whether all the training data conforms to a single or multiple pattern. If, say, two conflicting patterns of data can be separated, then they can be used for training two separate networks, each of which concentrates on its own group (they become 'local experts')"*. Adair, et al, (1996) also hypothesise that sub-markets can be identified by stratifying the market into increasingly homogeneous subsets. Clearly, the problem of representing location in mass appraisal models must be resolved.

In addition to the general appraisal research using neural networks, Collins and Evans (1994) attempted to model the effect of aircraft noise on residential property values. They conclude that neural networks exhibit *"powerful pattern recognition properties"* and can *"successfully discern complex value effects, including that of aircraft noise, in house price analysis"* (Collins and Evans, 1994).

Although most comparative studies of MRA and ANNs favoured the latter technique (Tay and Ho, 1992; Do and Grudnitski, 1992; Collins and Evans, 1994; Lam, 1996; McCluskey, 1996; Borst, 1991), ANNs are certainly not without criticism. Amongst

observations, Worzala, et al, (1995) reports a difference in results when using different ANN simulation packages; danger of 'overtraining'; trial and error approach required to set model parameters; model inconsistencies; and difficulty of use. The authors suggest *"extreme caution is necessary when applying the neural network technology to financial applications"* (Allen and Zumwalt, 1994 - in Worzala, et al, 1995), and repeat this warning for real estate appraisal (Allan and Zumwalt, 1994).

In a second paper (Lenk, et al, 1997), the authors suggest a lack of theory in the ANN field renders the fixing of model parameters, such as the number of hidden nodes and the training period, to the realm of 'trial and error'. Furthermore, the authors suggest this is the reason that ANNs are so often labelled as 'black-box' architectures. In a more sweeping statement, Lenk, et al, (1997) suggest that *"caution should be exercised and that more technical knowledge is needed before private and public confidence can be placed in these techniques for property valuation and lending decisions".*

McGreal, et al, (1998) also conducted a study of ANNs for the purpose of generating residential property appraisal models, concluding: *"Whilst some very close predictions are possible, others can deviate appreciably from the sale price. Under such circumstances the use of neural networks for mass appraisal purposes must remain problematic".* Although the authors do comment that for homogeneous data there is a *"tendency for better results."* (McGreal, et al, 1998)

Furthermore, there is a more pressing problem that undermines the acceptance of neural networks in the appraisal profession to date. A problem often termed as the 'black-box' syndrome. This in fact does not refer to the setting of model parameters as earlier suggested (Worzala, et al, 1995) but to the inability of ANN models to divulge their reasoning processes in a cognitive manner. Hypotheses learned by ANNs are difficult to comprehend as typically they consist of many real-value parameters. These parameters describe the relationship between the input features and the output value. Non-linear functions, represented by hidden units in a network, combine the input features thus allowing the model to take advantage of inter-dependencies within features (Craven, 1996).

The ability to furnish users with explanations of the reasoning process or underlying functionality is an important feature of any model (Clancy, 1983). Explanation facilities are required both for user acceptance and the validation of reasoning procedure (Davis, et al, 1977). In expert systems, explanations are typically provided by tracing the 'chain of inference' during the reasoning process (Southwick, 1991). This is a difficult task when analysing neural networks as they do not have explicit or *"declarative knowledge"* (Diederich, 1989).

2.6 Expert Systems

Another approach that has been considered, is the development of an Expert System (ES) that replicates the procedures taken by an expert human valuer. An ES is a computer system, which contains knowledge pertaining to an area of human specialisation. The system can implement that knowledge in such a fashion as to be able to act as a consultant expert in that field of specialisation (Gronow and Scott, 1985). Developing an ES capable of giving advice, based on a set of responses, requires *"a process of eliciting, interpreting and representing the knowledge for a given domain"* (Kidd, 1986).

A number of researchers considered ES to be a good basis for modelling residential (Gronow and Scott, 1987; Grant and McTear, 1992; Boyle, 1982; Jenkins, 1992) and commercial (Czernkowski, 1990; Nawawi, et al, 1996) property appraisal. In order to develop an expert valuation system, the knowledge possessed by an expert must be acquired or elicited to construct a set of domain rules. This process can take the form of interviews and questionnaires, from which domain rules are formulated by a knowledge engineer. Alternatively domain rules can be supplied by a team of expert valuers (Boyle, 1982).

This stage is fundamental to the development of an ES as the *"entire validity of expert systems depends crucially on the capture of the true nature of the procedure of the professional experts"* (Boyle, 1982). However, it is well documented that this stage is notoriously difficult to complete. In property valuation, *"the sale prices are the consequence of many expert and amateur judgements. The task of finding a unifying set of operational expert rules is not easy - given the conflicting opinions experts can have"* (Tay and Ho, 1992). Furthermore, valuation tends to be *"seen as an art not an exact science and valuers find it difficult to articulate the underlying*

process of their practice" (Grant and McTear, 1992). More worrying perhaps is that experts are sometimes unwilling to participate in lengthy knowledge acquisition exercises, possibly being anxious not to expose the existence of gaps in their knowledge or worse any incorrectness in their methods (Jackson, 1986).

A further hurdle restricting the development of an expert valuation system, is the problem of knowledge maintenance. ES do not *have "mechanisms to deal with any changes in their decision making environment"* (Goonatilake and Khebbal, 1995). They cannot learn *from "external changes in their operating environment or field of specialisation"* (Goonatilake and Khebbal, 1995). Successful ES are therefore restricted to very narrow domains under limited operating systems (Goonatilake and Khebbal, 1995). Due to their narrow domain and inability to react to change, ES are described as being *"Brittle"* in nature (Holland, 1987).

However, the advantages of using an ES as a valuation tool is clearly that *"they are constantly available, consistent in judgement, relatively cheap and have excellent memories. They can deal with large quantities of data when available yet can also offer advice in the absence of complete information"* (Gronow and Scott, 1987). Despite the brittleness of the technique, a valuation expert system modelled on properties from North Cardiff, predicted *"85% of these properties to within 0 and 2.5% of the value returned by the valuer. The remaining 15% came within 5%"* (Gronow and Scott, 1987). It is however interesting to observe that transaction price was revealed to this ES prior to its estimation. This perhaps in the light of more recent studies, highlighting the high correlation between transaction price and value (Gronow, et al, 1996), may account for this degree of accuracy.

Based on these results, Scott concludes it is possible to represent valuation expertise within the framework of an expert system. Furthermore, Jenkins (1992) concludes that: *"Expert system software of the 1980's combined with the hand held data collection tools of the 1990's will circumvent the obstacles to a rational valuation model".*

2.7 Summary

Before drawing conclusions from this literature review, it is worth summarising the main points covered:

The traditional DCC method has come under widespread scrutiny as a result of the financial deprivation caused by the house-price crash of the late 1980's. There is evidence (Almond, 1999) that some of the valuation profession has met this with an increased use of IT resources such as database technology to attempt to reconcile the lack of data to base judgements on given the 'thin market in traded properties' and the confidentiality of property transactions in England and Wales.

Some of the main criticisms resulting from this scrutiny have been the anchoring of judgement of value to only two or three comparable properties that are often out of context with respect to date or location. Similarly, research has shown that transaction price has an overbearing influence on the value that professional valuers provide - 90% of values returned by valuers were within 5% of the agreed and non-professionally attained transaction price (Gronow, et al, 1996). This perhaps reflects the habit of valuers in dispensing with the DCC adjustment grids and performing ad hoc adjustments (Almond, 1999).

Driven by the desire to move valuation from an art form to a science, alternative methods of valuation based on mathematical techniques using transaction data are being investigated. The most often researched method - Multiple Regression Analysis - relies on past transactions to set parameters determining the influence each measured attribute has on value. The method is simple and based on a well-defined and proven mathematical basis. However, it is limited to linear relationships and greatly effected by multi-colinearity and skewness. Linear programming, proposed by Wiltshaw (1991a,1991b,1993), is also constrained to models containing linear relationships. An alternative technique that overcomes this restriction comes from the field of artificial intelligence. Neural Networks are able to model non-linear relationships with ease, a functional form that is so intuitively present within property data. However, neural networks suffer from their own complex distributed nature - a form that makes explanation of functional form and reasoning somewhat difficult.

Expert systems and expert databases are also said to have a part to play in the scientific approach to valuation. Although, these approaches are dependent on eliciting knowledge from experts, a task which often runs aground given the non-

linear nature of expert knowledge, the transparency of knowledge representation provides the techniques with a firm foundation.

2.8 Conclusions

From this exposition of the published literature, it is possible to address a number of important questions, in particular:

Why is there such an interest in residential property appraisal?

Clearly, the interest in residential property research comes firstly from the consequences of the property crash in the late '80s and secondly from the desire of academia to increase the awareness within the valuation profession of new and emerging IT techniques.

What techniques are available?

The literature refers to a number of alternative techniques, of which the following were the most frequently mentioned:

- Multiple Regression Analysis (MRA)
- Linear Programming (LP)
- Artificial Neural Networks (ANNs)
- Expert Systems and Expert Database (ES and ED)

MRA is the most commonly used technique. However, the drawbacks of this techniques are its restriction to modelling only linear relationships and the danger of reducing modelling ability from unresolved skewness and multi-colinearity.

Linear programming is a technique proposed only by Wiltshaw (1991b) as a mathematical approach to modelling residential property values. This technique is based on firm mathematical foundations. However, this same mathematical structure poses a restriction on the application of this technique due to its algebraic requirements.

Next to MRA, artificial neural networks are the most commonly used technique within published literature. Their ability to easily model non-linear relationships - that are instantly recognisable within residential property data - puts this technique ahead of

MRA. However, the drawback of this technique is that the underlying functional form is less easy to interpret than the co-efficients produced in regression analysis due to its non-linear distributed structure.

In addition to these regression techniques, a further line of research is the development of an expert system that replicates the actions of the professional valuer. This is perhaps the easiest of techniques to comprehend but also the most difficult to build. The 'experience factor' used by professional valuers is difficult to replicate in explicit rule based structures. However, it could be said that appraisers impose linear constraints on the valuation problem to facilitate comprehension and it is this linear 'expert' interpretation that is modelled in such an expert system (Jenkins, 1992).

How effective are the alternative techniques?

To establish the effectiveness of the alternative techniques various benchmarks have been published and competitions held. However, the authors of most papers often cite restrictions that prevent such selections being made.

What challenges remain?

Despite a considerable quantity of high quality literature and research, there are still clearly a number of unresolved issues. Issues such as the need to establish a clear definition of the factors that influence value; the requirement to establish methods that are accurate beyond a tight well defined residential area; a strategy for combining the strengths of the investigated model into an holistic model; and, an appreciation from the valuation professionals as to the scope of alternative modelling techniques in everyday residential property valuation.

Empirical analysis suggests that neural networks *"seem particularly well suited to finding accurate solutions in an environment, such as residential appraisal"* (Do and Grudnitski, 1992). This is due to the fact that residential property data is *"characterised by complex, noisy, irrelevant, or partial information or imprecisely defined functional models"* (Do and Grudnitski, 1992). Furthermore, due to the comparative ease of data preparation and interpretation of results (Borst, 1991), neural networks overcome *"the methodological problems of multiple regression"* (Do and Grudnitski, 1992).

Despite the criticisms levelled at neural networks, results of various studies show *"that there are plenty of avenues available for future neural network research in real estate appraisal"* (Worzala, et al, 1995). Furthermore, *"continued research in this area is important and necessary before the final verdict on the use of neural networks in real estate appraisal can be decided"* (Worzala, et al, 1995).

There is an abundance of research potential in the field of residential property appraisal, with many unanswered questions relating to MRA and neural networks (Worzala, et al, 1995, Mackmin, 1994). Future research should concentrate on complementing the weakness of one method with the strength of another (Goonatilake and Khebbal, 1995) to provide the framework for a holistic model (Gronow, et al, 1996). However, it is noteworthy that the systems being developed are to act as instruments rather than appraisers. Researchers, in the most part, agree that *"Like an automatic pilot, the real decisions have to be taken by the professional who acts after the instruments have produced the basic information on current conditions"*. (James, 1994)

3. MODELLING HOMOGENEOUS DATA

This chapter describes empirical work designed to show the effectiveness of ANNs in predicting property values, using data selected from a homogeneous area. Alternative machine learning paradigms are also presented as useful methods of estimating council tax bands.

3.1 Introduction

The analysis of house price data to discover underlying value functions has for some time been performed using statistical techniques such as MRA (Adair and McGreal, 1986, Antwi, 1995). The objective is usually to establish a pattern that can be used to estimate the value of a previously unseen property or, to discern the effect a particular variable, such as noise pollution, has on property value.

Tay and Ho (1991) considered the use of Artificial Neural Networks (ANNs) for this purpose and reported a favourable performance compared with the more traditional MRA approach. Evans, et al, (1992), also investigated this approach in their study of properties in England. The purpose of this chapter is to mirror the work performed by Evans, et al, (1992), Tay and Ho (1992) and others to attempt to replicate these approaches using the training data that will be used for subsequent research.

The Chapter begins with an outline procedure for developing an ANN appraisal model. Following, a simple ANN solution is introduced together with empirical evidence. The approach is critically discussed followed by an analysis of the approach in respect of council tax appraisals.

3.2 Building an ANN Appraisal Model

The procedure for modelling residential property values using neural networks is as follows:

- obtain a database of past transactions;
- identify features within the data that impact on value;
- recode non-numeric data into a suitable numeric format;
- decide on an efficient neural network architecture;
- train network with a subset of data;
- validate network with a different subset of data;
- use network to predict the value of a previously unseen property.

Obtaining a good database of previous transactions is a non-trivial task. Gronow and Scott (1987) noted *"Lack of data is a fact of life throughout the property valuation world"*. Having obtained a database of previous transactions, it is then imperative that the right features are extracted. Jones (1996) *states "One of the major unresolved issues ... is how to adaptively select the input mapping for the neural network"*. Furthermore, Bishop (1994) states *"with finite sized datasets, reduction of the dimensionality may well lead to overall improvement in the performance of a classification system"*.

It might seem logical that the more information on each property, the better the model. However, this is not always the case, a large description for each property can sometimes confuse neural networks and MRA. Bellman (1961) calls this the *"curse of dimensionality ... beyond a certain point, adding new features actually leads to a reduction in the performance of the classification system"*.

3.3 Data Pre-processing

The data mining process (or knowledge discovery process) is the efficient discovery of valuable, non-obvious information from a large collection of data. This process starts with data preparation, next neural network models and architecture details need to be considered, followed by methods of training, testing and post-processing.

"There is always the question of if we have enough data. Then there is the issue of if we have clean, reliable data. Finally, and most importantly, is the determination of whether we have the right data. Only someone who understands the data and what it means, can select the right data for a data mining operation." (Bigus, 1996).

Most businesses have some form of database system. However, some of this data may contain inaccurate values, missing data or other inconsistencies. Some form of data cleansing must be used in this case. This process may involve many different techniques including the use of application software or rule-based techniques, which evaluate each data item against meta-knowledge for the domain process, or visualisation techniques to highlight obvious outliers, or statistical techniques to deal with missing or incorrect values.

The next step is to determine what data is important for the modelling task. This step requires some detailed knowledge of the domain process. Having decided which fields are important in modelling the process, often this data needs to be put into a format that a neural network can use. This sometimes involves creating many fields to describe the data in an existing field, summarising many fields in one new field, creating ratios based on two or more fields or converting symbolic values into numeric values. This is usually achieved using database queries or spreadsheet analysis, but more recent neural network packages such as PREDICT include front end processing as an additional module. There are many methods for recoding symbolic data of which the following are most common:

> One-of-N codes - Also sometimes called dummy variables or binary codes, this method involves creating a field for each class in a particular variable. An example variable where this might be appropriate is heating type containing the classes or values {Full, Partial, None}. Each of the three separate fields for this example would contain a '1' if the heating attribute for that record matched the field name or '0' otherwise. Hence, a record with no heating would be recoded as {0,0,1}. The problem with this method is that it can be very costly in terms of network size, a single field may need 50 new fields to represent all its values.

Thermometer codes - Also known as continuous coding, this method is used when a definite ranking is known for each of the values needing recoding. Using the same example - and assuming a property is worth more if it has partial heating than if it has none, and worth even more if it has full heating, - then recoding could take the format: Full - 3, Partial - 2 and None - 1. Caution must be exercised when using this method, as an incorrect ranking will confuse the neural network.

Continuous variables such as number of bedrooms do not require any recoding, although benefit from being scaled or normalised. This translation sets the range of the data within the domain of the transfer functions used in the learning process.

Care must be taken when performing the data analysis stage, as poor data representation may result in a large network that has poor generalisation performance. A smaller network that includes good data representation will be able to generalise better, give better predictions and be easier to understand.

3.4 Neural Network Architecture

In addition to establishing an optimum data representation, it is also important to select suitable neural network architecture. The options available include the type of network, i.e. supervised or unsupervised; the size of the network, and the transfer function employed. The basic building block of the most commonly applied network for regression type problems, the multi-layered perceptron network, is the perceptron.

Basic Perceptron

The basic perceptron consists of a set of input nodes, a set of weights, a transfer function and one or more output nodes. Figure 3.1 is a schema of a basic perceptron.

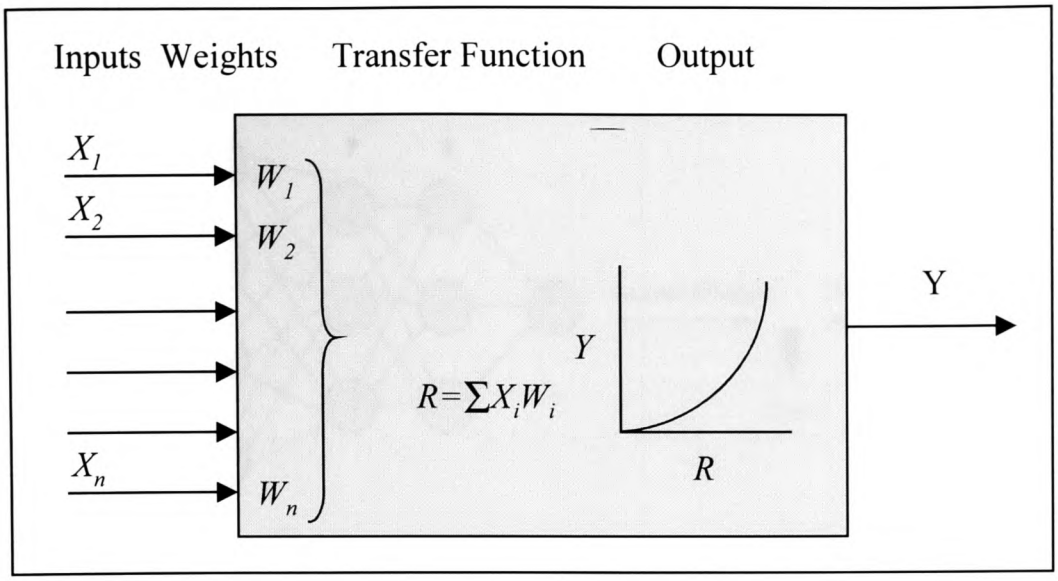


Figure 3.1 - Schematic of a Basic Perceptron

Where X_i is the input values and W_i the weights assigned to individual input nodes. A transformation function is used to generate Y (the output) from the weighted sum R. The transformation function can take many forms of which the most common is the regular sigmoid transformation function.

Training involves comparing the generated output with the desired output and adjusting the weights accordingly to reduce the global error across a set of training data.

A single perceptron is however unable to discern patterns that are not linearly separable and hence a more complex architecture such as the multi-layered perceptron architecture is required.

Multi-Layered Perceptron

The multi-layered perceptron (MLP) architecture is quite simply a network of perceptrons as shown in Figure 3.2.

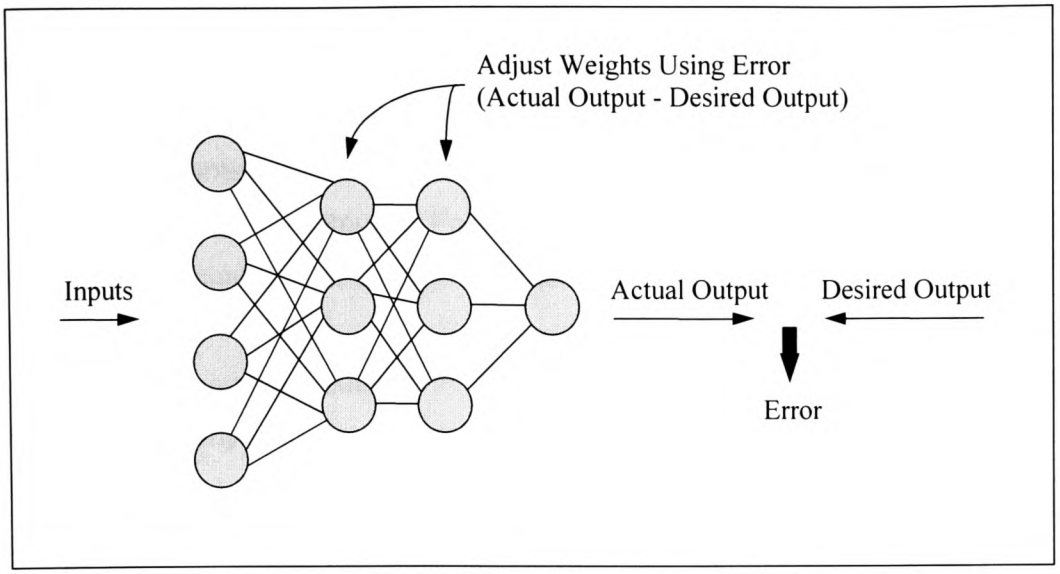


Figure 3.2- A Simple Multi-Layered Perceptron (feed forward back propagation) Network.

Determining the number of hidden nodes often involves trial and error, however there are some mathematical expressions that help. For example Zurada (1992) suggests the following equation can be used to evaluate the number of hidden nodes required in a back propagation network:

$$\textit{Number of Hidden nodes} = Log_2 \textit{ Number of output Classes}$$
$$\textit{where Number of Hidden nodes } <= \textit{ dimensions of pattern space}$$

However, alternative and in some cases contradictory methods have also been suggested. For instance, Eberhart and Dobbins (1990) concluded *"There are no theoretical guidelines for determining the number of hidden nodes to use on a given application."* The choice of structure becomes intuitive in nature based primarily on past experience. Nonetheless, the importance of this decision must be stressed, as a large number of hidden nodes can significantly increase processing time and reduce the generalisation property of the neural network (Fahlman and Lebiere, 1988).

The final step in the data preparation stage is the division of the source data into three data sets. One data-set is used to train the network, one to iteratively test the network during training and a final data-set to assess the accuracy of the neural network model - these data-sets are normally called the Training set, Validation set and Testing or Hold-out set respectively. Normally this division of the source data-set

is done randomly with approximately 80% of the data in the training set, 10% in the test set and 10% in the validation set. However, there may be times when this random selection method is not representative of the whole process. For example, if the source set contains 100 cases where a process resulted in an output of True and only 10 resulting in an output of False, then to be representative, the training set should contain equal numbers of both outcomes.

Finally, the inclusion of outliers in the training set can dramatically reduce the effectiveness of any modelling technique. This is also true for neural networks.

3.5 Empirical Evidence

This section details the empirical work undertaken to establish whether neural networks can assist in the residential property appraisal process. The work makes use of a database containing details of residential property transactions handled by a UK building society, in the Cardiff area. The database is a subset of a much larger database that contains residential property appraisal data for the whole of the United Kingdom for the period January 1993 to December 1995. While it is believed that the conclusions reached are generic to many residential property appraisal situations, the results are specific to the data set used.

A database as described in Appendix 1.1 was obtained from a leading lending institute in the UK. From this, data from the Cardiff area was selected as this facilitated access to local a priori knowledge. The area contains a broad range of wealth scales and house types, and is, therefore, heterogeneous in nature.

Evans, et al, (1992) investigated claims (Tay and Ho, 1992) that neural networks could provide a successful means of analysing residential property price data. Their research considered 34 post 1960 houses and bungalows in 14 different streets in a Midlands town. An average difference between the predicted value and the transaction price of between 5 and 7 per cent was reported. They concluded that *"the technique [would be] highly suitable for applications such as taxation valuation ...[or]... an additional tool to speed the normal valuation process ...[and] ... for preliminary or so-called desktop valuation before inspection of the property"*.

Their results compare very favourably with those obtained using the building society database. Using the building society database, the mean absolute percentage difference between the value returned by the valuer with the value predicted by the neural network is approximately 28%. Further tests using the building society database led to the conclusion that the variance within the database was too great to enable a single back propagation network to model the database accurately. (This was also true for MRA techniques that achieved a mean absolute error of 27%.)

One reason for such a high error rate is the heterogeneity of the data set. However, it was envisaged that the error rate would decrease by employing a number of different techniques. The development and application of these techniques - listed below - formed the basis of the main empirical work:

- Identifying and modelling homogeneous sub-sections of the data (as opposed to modelling the whole heterogeneous data set)
- Imputing a priori knowledge pertaining to a ranking for location
- Adding to the depth of the database (using new information or by considering averages within the data set)

As a first step, confirmation that a neural network could model properties from a homogeneous area described using the building society features was required. The area selected was Roath (a district of Cardiff). The training and test files contained 37 and 13 records respectively - after removal of non-typical records (often referred to as outliers) identified by inspection of attribute histograms (Attributes used in this analysis are indicated in Appendix 1.1[2]). Table 3.1 shows the results obtained using the test file. The analysis shows a mean absolute percentage error of 7.5%, reducing to 5.9% after additional outlier removal (based on the mean result for a number of trials). This is comparable with the results reported by Evans, et al,. (1992). This work supports the claims made by Evans, et al,. (1992) and Borst (1991) that a neural network can model an appraisal function (albeit for a homogeneous area).

[2] For the remainder of the empirical work, the same housing attributes were used as used in this study (see Appendix 1.1). Additional attributes drawn from the 1991 UK Census are used in the 'Stratification' work (Chapters 5, 6 & 7). The particular attributes used for each study are indicated in Appendix 1.2.

Table 3.1- Results Obtained for the 'Roath' Test Set.

Record Number	Value	% Absolute Error
1	£34,250	10
2	£52,500	8
3	£47,000	11
4	£49,950	11
5	£42,000	3
6	£41,000	17
7	£45,000	6
8	£55,000	7
9	£65,500	4
10	£54,000	10
11	£61,000	4
12	£42,000	6
13	£36,000	1

From the results it can be seen that the neural network is able to model the data in the training set with a degree of generalisation that allows estimation of target values for the records in the test set with a mean absolute percentage error of only 7.5%.

The success achieved for the homogeneous training data was not however repeated in studies using ANNs and MRA for heterogeneous data-sets. The results obtained did not approach any acceptable accuracy targets with some of the worst predictions being over 100% greater or less than the actual sale price of the property. Clearly, the homogeneity of the data plays a significant part in the accuracy of the appraisal model.

This conclusion was for the most-part known at the outset of the research and indeed one of the aims of the research was to investigate methods of modelling heterogeneous data-sets given that single-model attempts using the whole data-set will often result in failure. However, before commencing that part of the research it was considered profitable to use the fact that it was possible to create appraisal models based on homogeneous data and to investigate how such models could be created for the purpose of property tax valuations.

3.6 Property Taxation

One useful application of a computerised residential appraisal model is the valuation of properties for taxation purposes, and in particular - in the UK - the council tax. Unlike individual valuations required for mortgage purposes, council tax valuations are perhaps more suited to computer assisted mass appraisal where valuation accuracy can be stated within a given tolerance. Furthermore, local governments

have to deal with valuation of all real estate within their boundaries within a short space of time and within limited budgets.

In the UK, property tax for residential properties is levelled according to a banding system based on the taxation brackets as shown in Table 3.2.

Table 3.2 - Breakdown of Council Tax Brackets in the UK (1997/1998)

Taxation Band	Lower Limit (£'000s)			Upper Limit (£'000s)
A			⩽	40
B	>	40	⩽	52
C	>	52	⩽	68
D	>	68	⩽	88
E	>	88	⩽	120
F	>	120	⩽	160
G	>	160	⩽	320
H	>	320		

A sliding scale is used to tax the occupiers of each residential property according to the band in which their property resides. Revenue raised from this tax is used locally to fund refuse collection; environmental health; cemeteries; recreation; car parking; street cleansing; housing/council tax benefits; and other services and improvements.

The obvious difference between mortgage valuation and council tax valuation is that the target output is no longer a point (e.g. £25,500) but a class (e.g. Band A). This allows the appraisal model to be built as a classification one rather than regression one. Neural networks are as well suited - if not more suited - to classification problems, and there are also alternative 'Artificial Intelligence' techniques that could be employed such as tree induction algorithms. To provide an estimate of the potential such techniques have as part of a Computer Assisted Mass Appraisal (CAMA) system for council tax valuation, three techniques will be employed: neural network; MRA and Tree Induction (Quinlan, 1986). These methods are described in the next two sections, with specific configurations for classification problems noted.

Configuring an MLP for Classification Problems

To enable a multi-layered perceptron to model classification problems, the number of output nodes needs to be increased from a single node to multiple nodes - one for each target classification. Figure 3.3 illustrates the architecture required to model a classification problem having 6 input features and 4 output classifications.

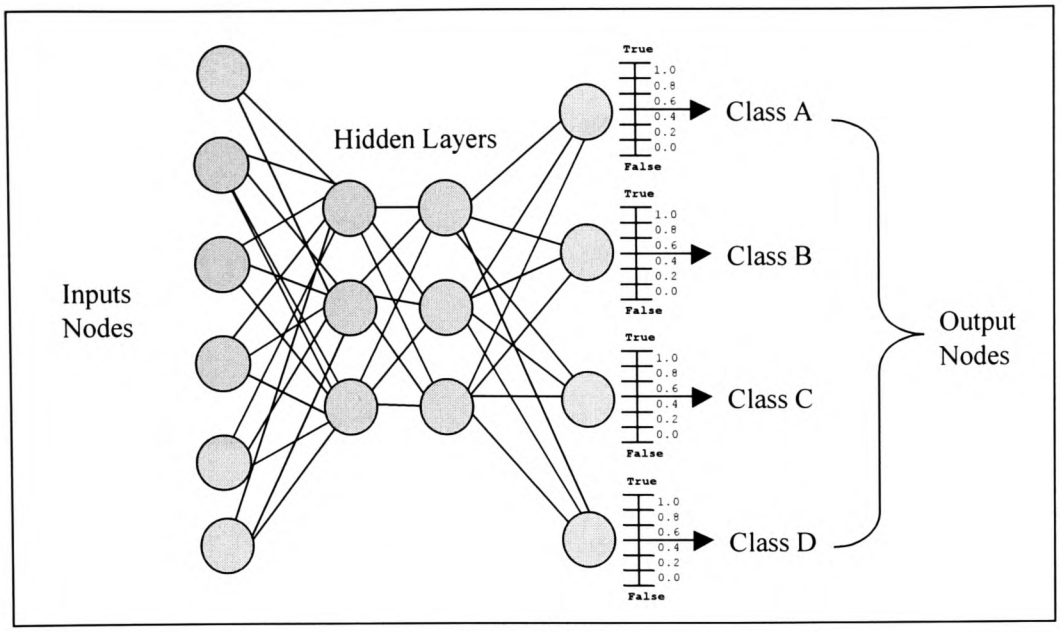


Figure 3.3 - MLP Architecture for Classification Problems

Classification can be achieved by assigning membership according to the magnitude of the output nodes. For example, with output values of (0.8,0.3,0.1,0.4) for output nodes 1,2,3 and 4 respectively, the input vector would be classified as 'Class A' as this has the highest 'truth' value.

The underlying functionality of perceptrons using Sigmoidal transfer functions tend to improved accuracy for binary classification problems, as a higher proportion of the domain of the transfer function maps to the upper and lower limits of its range (See Figure3.4).

For the purposes of modelling property taxation banding, one output node will be used to represent each council taxation band.

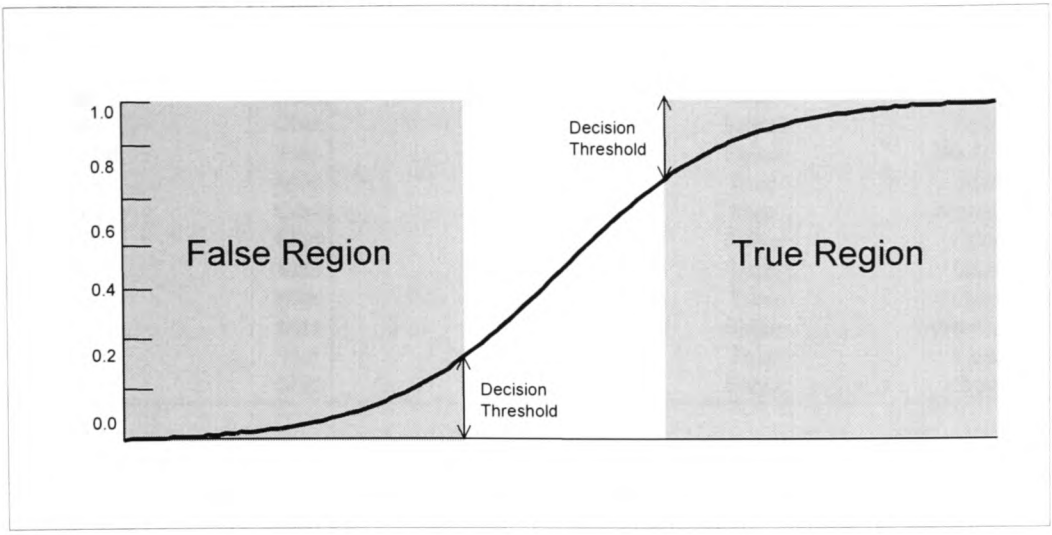


Figure 3.4 - Example of the Classification Bias of the Sigmoidal Transfer Function

Tree Induction Techniques

Tree induction techniques or decision trees, are an alternative to neural networks often used to model classification problems. The most common implementation of tree induction techniques is Quinlan's C4.5 and C5.0 (Quinlan, 1993). Tree induction uses a divide-and-conquer approach to growing decision trees. Quinlan (1986) describes the concept of decision trees as:

> If all training cases belong to a single class then
> The tree is a leaf labelled with that class.
> Otherwise,
> Select a test, based on one attribute, with mutually exclusive outcomes;
> Divide the training set into subsets, each corresponding to one outcome; and
> Apply the same procedure to each subset

Once constructed, such a decision tree can be used to classify a new, unseen case described in terms of the same attributes. Starting at the root of the tree, if the current node is a leaf the case is assigned the classification of that leaf. Otherwise, the outcome of the test at that node is determined and the corresponding branch is followed.

Table 3.3 provides an example training set for the tree induction approach. The goal of the training process is to determine whether a coat should be worn for various weather conditions.

Table 3.3- A Small Training Set

Outlook	Temperature	Humidity	Windy	Class
Sunny	Hot	High	False	No Coat

Sunny	Hot	High	True	No Coat
Overcast	Hot	High	False	No Coat
Rain	Mild	High	True	Coat
Rain	Cool	High	True	Coat
Rain	Cool	Normal	False	Coat
Overcast	Hot	Normal	False	No Coat
Sunny	Mild	Normal	True	Coat
Sunny	Cool	High	True	No Coat
Rain	Mild	Normal	False	Coat
Sunny	Mild	Normal	True	Coat
Overcast	Mild	Normal	True	Coat
Overcast	Mild	High	False	No Coat
Rain	Hot	Normal	True	Coat
Rain	Mild	High	False	Coat

Derivation of a decision tree requires a splitting criterion to be established, generally this focuses on class entropy - branches are investigated in the order to which they increase the homogeneity of generated subsets. Figure 3.5 provides an illustration of a simple decision tree that could be induced from the above training set.

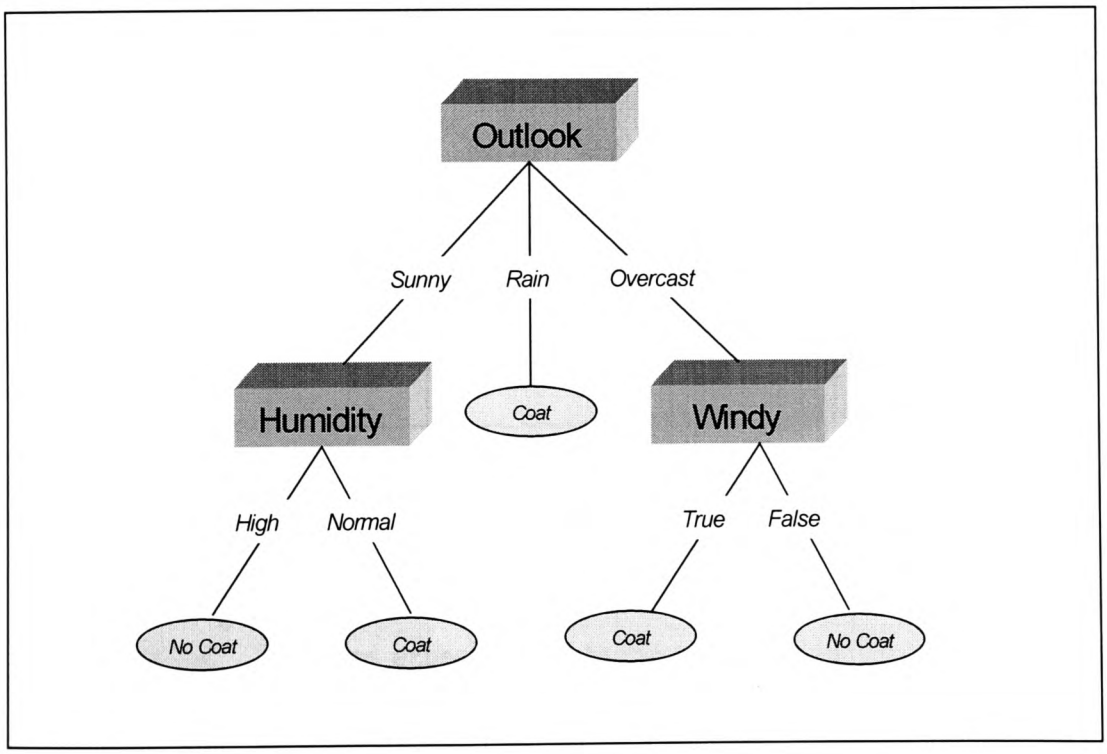


Figure 3.5 - A Simple Decision Tree

If the attributes are adequate - that is, duplicate input vectors are members of the same output class - then it is always possible to induce a decision tree from the training data.

Empirical Analysis

In order to assess the potential of tree induction, MRA and ANN techniques as methods for modelling council tax bands of residential properties, a slice of mortgage transaction data is converted so to express as its output the council tax band for which its mortgage value resides.

Results

The results of this analysis are presented in Table 3.4. The individual percentages reflect the levels of accuracy for each technique in predicting values within the respective bands. For example, for the homes valued in Band A in the training set, the ANN model correctly predicted 93% of them as being Band A properties, and therefore 7% were incorrectly predicted as having a value outside the Band A range.

Table 3.4 - Results of Council Tax Banding Analysis

	Band A	Band B	Band C	Band D	Band E	Band F	Band G	Band H	Average
ANN Train	93%	60%	65%	54%	90%	67%	0%	N/A	80%
ANN Test	75%	50%	60%	0%	94%	N/A	N/A	N/A	65%
C5.0 Train	100%	70%	77%	75%	96%	100%	N/A	N/A	92%
C5.0 Test	94%	7.5%	18%	46%	25%	N/A	N/A	N/A	59%
MRA Train	86%	54%	72%	28%	60%	54%	50%	N/A	57%
MRA Test	45%	78%	23%	90%	25%	N/A	N/A	N/A	52%

From the results it can be seen that the ANN implementation out-performed the Tree induction C5.0 implementation and the MRA model for the test set. Also, it is notable that the C5.0 trials fared considerably better when predicting the council tax band for the training set. This is as expected due to the fact that tree induction techniques learn examples within the training set compared with ANN models that learn general underlying patterns.

However, the notable advantage tree induction techniques have over ANNs is their ability to explicitly divulge the structure of the generated model. Specifically, tree induction techniques can be translated directly into process rules.

Rules Induced by Quinlan's C5.0 Algorithm

The following rules were extracted using Quinlan's C5.0 algorithm for the analysis described earlier (results quoted in Table 4 of this Chapter). The rules can be read as 'expert system' rules with the decimal shown in brackets at the end of each rule being a confidence factor based on the sample size satisfying each rule.

Band A properties are valued between £0 and £40,000

If House_Type is 'Mid-Terraced House'
And Floor_Area <= 170
Then Council_Tax_Band is Band A (0.833)

IF Floor_Area < 142
Then Council_Tax_Band is Band A (0.580)

If Date_Built <= 1975
And Floor_Area <= 142
Then Council_Tax_Band is Band A (0.809)

Band B properties are valued between £40,000 and £52,000

If House_Type is 'Mid-Terraced House'
And Floor_Area > 170
And Condition is 'Fault'
Then Council_Tax_Band is Band B (0.750)

If House_Type is 'Semi_Detached House'
And Date_Built < = 1975
And Floor_Area <= 137
Then Council_Tax_Band is Band B (0.583)

Band C properties are valued between £52,000 and £68,000

If House_Type is 'Semi-Detached'
And Bedrooms <= 3
And Date_Built > 1929
And Floor_Area is between 116 and 137
Then Council_Tax_Band is Band C (0.833)

If House_Type is 'End-Terraced House'
And Date_Built > 1979
Then Council_Tax_Band is Band C (0.750)
If House_Type is 'Semi-Detached House'
And Floor_Area is between 137 and 142
Then Council_Tax_Band is Band C (0.667)

Band D properties are valued between £68,000 and £88,000

If House_Type is 'Semi-Detached House'
And Date_Built > 1975
Then Council_Tax_Band is Band D (0.750)

If House_Type is 'Semi-Detached House'
And Bedrooms <= 4
And Floor_Area > 142
And Condition is 'No Fault'
Then Council_Tax_Band is Band D (0.583)

If House_Type is 'Semi-Detached House'
And Date_Built > 1914
And Floor_Area > 142
Then Council_Tax_Band is Band D (0.750)

Band E properties are valued between £88,000 and £120,000

If Bedrooms > 3
And Garage is 'Double Garage'
And Floor_Area <=166.5
Then Council_Tax_Band is Band E (0.833)
If House_Type is 'Detached House'
And Garage is 'Single Garage'
Then Council_Tax_Band is Band E (0.536)

If House_Type is 'Semi-Detached House'
And Bedrooms > 4
And Date_Built < 1914
Then Council_Tax_Band is Band E (0.750)

Band F properties are valued between £120,000 and £160,000

If House_Type is 'Detached House'
And Bedrooms > 3
And Condition is 'Fault'
Then Council_Tax_Band is Band F (0.800)
If House_Type is 'Detached House'
And Garage is 'Parking Space'
Then Council_Tax_Band is Band F (0.625)

If Bedrooms > 3
And Garage is 'Double Garage'
And Floor_Area > 166.5
Then Council_Tax_Band is Band F (0.667)

3.7 Summary and Conclusions

This chapter began with a 'best practice' for preparing property data to form a training set for a neural networks. The main issue being the representation of non-numeric data in a form suitable as a neural network input.

Following, a simple approach to building a neural network appraisal model was detailed and proven using training data selected from a homogeneous region. The aim of this exercise was to assess the validity of the data to act as a training set for homogeneous analysis prior to pursuing research into modelling data selected from an heterogeneous area. Validity was measured by comparing models trained using the data against published benchmarks. The data was shown to be useful as training set for a neural network model given the selected data came from a relatively homogeneous area.

An application for this simple approach is suggested as being the estimation of property value for taxation purposes. As in the previous example, a neural network model was trained with a single output node representing value, this was compared with a further neural network model representing value as a banded classification. An additional model based on Quinlan's tree induction techniques was also investigated. The results show that Quinlan's model performed significantly better than the neural network models when predicting examples from the training set. However, it was the neural network models that came to the forefront when predicting previously unseen examples due to their generalisation qualities.

Overall, it can be concluded that neural networks are quite able to represent valuation functions buried within transaction data selected from homogeneous regions. The question remaining being: Is this quality also shown for models trained using data from heterogeneous regions or will further technique development or data enhancement be required?

4. LOCATION, LOCATION AND LOCATION

This chapter highlights the influence location has on the value of a residential property. Methods of including locational descriptions - in particular measures of neighbourhood quality are investigated.

4.1 Introduction

The first aspect of a property to be considered by 'house-hunters' is usually the location. Property has a fixed location. Therefore a property's value will depend on the benefits of owning a property at that specific location. These benefits, or put another way, the enjoyment of a property, will depend upon the general environmental factors and specific local factors, the relationship between employment opportunities, communications and the general facilities of an area. (Mackmin, 1994)

The nature of the neighbourhood and the immediate surrounding properties are crucial factors influencing a buyer's perception of a house. A socially deprived or underprivileged area will display that fact in the deterioration of the urban fabric, including the deterioration of the physical condition of homes (Mackmin, 1994). The levels of crime and vandalism are also indicative of an area's desirability. Proximity to roads - particularly motorways, railways, rivers, village greens, sports fields, parks etc. - may give rise to higher or lower property prices. For example, some riverside properties are highly desirable while others form the lower end of the market.

Mackmin states: *"The general economic climate together with the quality of different residential areas creates a pattern of values for a defined market."* (Mackmin, 1994)

It is perhaps not surprising then that so often the primary influences on property value are often quoted as Location, Location and Location. Boyle (1982) explains that *"two otherwise identical houses, separated by a few miles, can vary in price by as much*

as 50 per cent". He suggests *that "any model must give location an extensive treatment"* (Boyle, 1982).

Representing location as a numerical surrogate is non-trivial. To gain a full comprehension of the quality of an area - and hence those things to consider when valuing a residential property - takes time and experience. Moreover, the 'quality' of an area is a metric that is far from static. Classic examples of this dynamism being the decline of cities such as Detroit with the demise of its car manufacturing industry and the dramatic decline of industrial South Wales with the sudden closing of the coal mining industry (Mackmin, 1994).

4.2 Constructing a 'Base Value'

One method of ranking location is to consider average house value in each location. Jenkins (1992) used this method as a base value in a heuristic approach to valuation. In this study, the average house value for each district represented location. This approach was investigated by the Author using the Cardiff database, with Table 4.1 defining the averages extracted and the results of training and testing an MLP network on all 1321 records.

Table 4.1 - Results of Adding Average Values as Inputs to MLP Model

Data-set	Mean absolute % error
Cardiff Data set.	27%
Mean value by district.	26%
Median value by district.	26.2%
Mean & Median value by district.	27.7%
Mean value by type by district.	24.49%
Median value by type by district.	24.49%
Mean & Median value by type by district.	24.80%
Mean value by bedrooms by district.	24.76%
Median value by bedrooms by district.	24.86%
Mean & Median value by bedrooms by district.	24.95%

Using average house price as a surrogate for location increased the accuracy of the model. The results show that average values for house type, within district, gives some representation of location. However, although some improvement in accuracy is observed, it is imperative that more depth is added to the database from other sources. Clearly, even with average values used as features, the variance within the data set is still outside any acceptable threshold. However, this approach, where an appreciation for average prices in a location is developed, is one used by many buyers whom, arguably become amateur valuers for a short while.

4.3 Geodemographic Indicators

Neighbourhood quality could be estimated by considering geodemographic indicators that band neighbourhoods into a set of discrete classifications. A customer targeting tool called ACORN™, developed from Census analysis, attempts to provide a *"powerful tool"* (CACI, 1996) to *"address the complexity of consumer markets"* (CACI, 1996). Indeed, the Nationwide Building Society's house price index uses the ACORN™ classification system.

Currently, there are 5 major companies in the UK that provide geodemographic systems. Table 4.2 details the major geodemographic systems.

Table 4.2 - Geodemographic Classification Systems based on 1991 Census

Classification System	Supplier	Non-Census Data Used
ACORN91	CACI	No
MOSAIC91	CCN	Yes
DEFINE91	Infolink	Yes
IMAGES	Equifax Europe Ltd.	Yes
Neighbours & PROSPECTS	Euro Direct dB. Marketing Ltd.	No

These systems are used frequently in direct marketing as they allow companies to target customers using the assumption that behaviour patterns and expectations can be generally related to neighbourhood characteristics. (Sleight, 1993)

Commercial geodemographic indication data is however expensive, which precludes them from inclusion in academic studies. A further geodemographic system was developed as part of an ESRC research project and is now available to academics. This system, known as GB-Profiles '91, classifies EDs in England and Wales and OAs in Scotland into a relatively small number of distinct residential area types (Openshaw and Wymer, 1994).

4.4 Census Data

The 1991 Census provides researchers and government with the *"most authoritative social accounting of people and housing in Britain"* (Dale and Marsh, 1993). Comparable statistics are generated for very fine geographical areas, the smallest of which being an enumeration district (ED) in England and Wales, and an output area (OA) in Scotland. The raw data collected from the Census can be cross-classified to provide *"powerful statistical insights into the social conditions of the population and its*

housing" (Dale and Marsh, 1993). Due to the inclusion of postcode in the 1991 Census questionnaire, it is possible to link social and economic data with housing stock data via a postcode to enumeration district table. Figure 4.1 shows the relationship between the abstraction levels for which Census data is available.

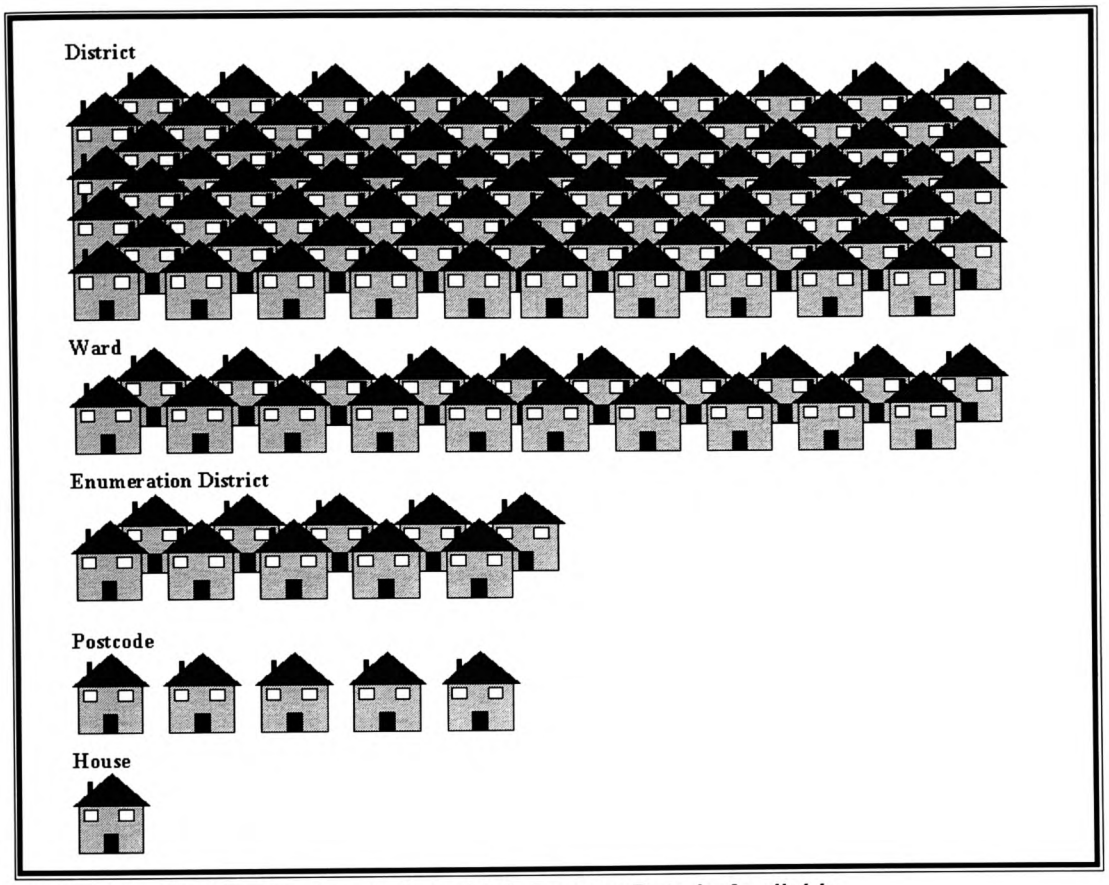


Figure 4.1 - Abstraction Levels for which Census Data is Available.

Ideally, to achieve a representative quality from the Census data, information should be extracted at the ED level (Note: all postcodes in an ED share the same Census characteristics). However, this is a labour intensive process, is best pre-empted with a study to determine which Census features are important from amongst the 20,000 available. Current methods of determining the features that impact on a dependent variable - in both neural networks and MRA - rely heavily upon a priori knowledge. Mathematical techniques exist, examples including principle component analysis and stepwise regression. However, these tend to be linear in nature and may result in a loss of information. Hence the former will be used as a first step and then mathematical methods used where appropriate.

4.5 Selecting Census Variables for Residential Appraisal Models

The statistics available from the 1991 Census represent the OPCS perception of need after an extensive consultation process. Questions were asked on households (amenities, tenure etc.) and individuals (age, ethnicity, occupation etc.) and these have been recombined under eight general headings: Demographic; Ethnic; Housing; Household Composition; Socio-economic; Migration; Health; Travel-to-Work.

Openshaw and Wymer (1994) provides a detailed 'audit trail' of the process of selection of variables for inclusion in geodemographic indicators, this approach is presented below for inclusion of Census variables in residential appraisal models.

Age Structure: Age gives a measure of the life cycle of individuals. Residential areas can sometimes be characterised by the age of the residents. For example, an area may have a large population of pensioners or pre-school infant etc.
Sex and Marital Status: The traditional married couple model is perhaps no longer dominant. Increasingly, households contain single people; students; cohabiting couples; working women and lone parents. These statistics may prove useful in describing the social make up of a neighbourhood and hence may impinge on value. Included in this section are statistics describing the number of working women (single; in couples; married) and lone parents.
Students: Students are increasingly forming a large proportion of the UK population. Housing in regions predominated by students are often owned by landlords (privately or in university owned blocks) and usually get divided into low cost rented bed-sits. When non-student housing is put on the market in such an area, buyers are often plentiful as they regard such purchases as an investment. It is not difficult to appreciate that the housing market in areas containing a high proportion of students operates differently from other regions.
Ethnicity: The ethnic make up of an area may have an influence on property value. However, Openshaw and Wymer (1994) warns of the danger of associating multi-ethnic areas with poorer areas and suggests that tenure is used to distinguish between *"financial stable and financially stressed multi-ethnic regions"* (Openshaw and Wymer, 1994).
Migration: Migrational statistics can help to provide a crude measure of neighbourhood quality. Inward migration may be attributed to investment (employment; new housing estates; leisure facilities) and also when summarised as pensioner migrants may indicate the attractiveness of an area to pensioners (e.g. coastal areas and retirement homes). Outward migration may be seen as the converse.
Tenure: Statistics describing nine tenures are available: Owner Occupied Outright; Mortgaged; Privately Rented (Furnished); Privately rented (Unfurnished); Rented with Business; Local Authority Rented; Housing Association Rented; Armed Forces Rented. Openshaw and Wymer (1994) states *"Generally, the rented category and especially accommodation rented from Local Authorities etc. has been used by researchers as a measure of lack of resources and residential insecurity. In Contrast, because of the financial commitment required to purchase a house, house ownership is seen as a surrogate for long term financial stability."*
Household Type: The general mix of housing stock in a region may have an impact on property prices. Statistics are available on detached; semi-detached; terraced; flats; bed-sits.
Amenities: The level of amenities enjoyed by households in an area can directly influence the value of a property. However, the number of properties where the WC or shower is shared is negligible.
Car Availability: Openshaw and Wymer (1994) used the percentage of households lacking a car as a measure of short-term financial deprivation. Statistics are available on those households with: no car; 1 car; 2 cars; 3+ cars.
Overcrowding: Deprivation can also be characterised by overcrowding. Openshaw and Wymer (1994) states that a value exceeding 1.5 persons per room indicates an overcrowded household. Housing size is also a useful measure, as larger houses require a greater income to maintain.
Household Composition: The composition of households in an area can also give an indication of the levels of income. Statistics are available describing: family type (couple with children aged 16-24; couple without children aged 16-24 etc.). The numbers of dependants are also available.
Socio-Economic: Census variables describing the proportion of people in an area who are economically active (employers, self-employed, employees etc.), their skills in an industry (managerial,

professional, semi-skilled etc.) and also the type of industry (agricultural, manufacturing etc.) are available. Statistics describing workers qualifications (higher degrees, diplomas etc.) are also available. Finally, the proportion of unemployed people in a community can be estimated using Census statistics.

Health: Limiting long term illnesses and the number of people in hospital accommodation can be estimated using Census statistics. However, general health issues are unlikely to have a marked effect on property prices.

Travel-to-Work: Clearly, the opportunity to live relatively close to your place of work is an advantage. However, it is also true that the industrial hub of a town or city is often surrounded by properties at the lower end of the market.

To investigate the usefulness of Census data, Census variables at the district level and the Enumeration District level were selected. A description of the selected variables appears in Table 4.3.

Table 4.3 - Census Variables Used in Analysis

Socio-Economic Group	
Employers and Managers (Large est.)	Employers and Managers (small est.)
Professional workers (self-employed)	Professional workers (employees)
Ancillary workers and Artists	Foreman and Supervisors (non-manual)
Junior non-manual workers	Personal Services workers
Foreman and Supervisors (manual)	Skilled Manual workers
Semi-Skilled Manual workers	Unskilled Manual workers
Members of Armed Forces	
Employment	
full-time Employment	On Government Scheme
part-time Employment	Unemployed
Self Employed	
Qualifications	
Qualified Persons	Higher Degree
Degree	Diploma
Qualified and on Government Scheme	Qualified and Unemployed
Age Ranges of Qualified Persons	
Housing Stock	
Detached Properties	Purpose-Built Flats
Semi-Detached Properties	Converted Flats
Terraced Properties	Bedsits
Tenure	
Owner Occupied (Outright)	Owner Occupied (Buying)
Privately Rented (Furnished)	Privately Rented (Unfurnished)
Rented from Housing Association	Rented from Local Authority
Amenities	
Shared Use of WC	Exclusive Use of WC
Central Heating	
Availability of a Car	
Households with no car	Households with 1 car
Households with 2 cars	Households with 3+ cars
Ethnicity	
White	Black Caribbean
Black African	Black Other
Indian	Pakistani
Bangladeshi	Chinese
Asian	Persons born in Ireland
Miscellaneous Variables	
Working Mothers (Part-Time)	Working Mothers (Full-Time)
Lifestages (age ranges of residents)	Overcrowding (persons per household)
Travel to work estimates	

Census data grouped at a district level contains on average 2816 people in 1528 households. At this aggregate level, many of the characteristics that are found in smaller groupings such as enumeration districts and postcodes are hidden (Dale and Marsh, 1993). However, it is useful to investigate whether high-level data imputation can aid the domain modelling process. Table 4.4 shows the results for a number of district level trials.

Table 4.4 - Results obtained when Census data at district level was used.

Description	Mean absolute % error
Control case (Whole Cardiff Data-set).	28.33
All Census Information (As Appendix 1.2)	26.36
Employment statistics.	26.89
People to Car Statistics.	27.31
Occupation statistics.	26.91
Ethnic groupings statistics.	27.94
Tenure statistics.	26.73
Property Type statistics.	27.24
Amenities statistics.	27.02

The results show an improvement in accuracy at this very high level of abstraction. The best performance, a percentage accuracy increase of 2%, was achieved when all the selected Census variables were added to the Cardiff data set.

Table 4.5 shows the results of a similar exercise performed using data extracted at the ED level, using the same property and Census variables. Here the mean absolute percentage difference between the predicted value using Census data and the value returned by the valuer was 7% closer than that achieved in the absence of Census data.

Table 4.5 - Results obtained for ED level analysis using all selected Census attributes.

Data Sample	Mean absolute % error
Cardiff Test set	20.23
Cardiff Test set & ED level Census statistics	13.54

4.6 Discussion

It is evident, from the results obtained that considerable improvement in accuracy is gained when Census data is added to property data for residential property appraisal modelling. This suggests that the influence of location on residential property values can be partially described using Census data. Along with the marginal increase in accuracy gained using average property values, it may be concluded that computer assisted appraisal models can be improved by including reference to locational and

demand side variables. However, the coupling of Census data with property data is a somewhat simplified approach, ignoring both the existence of sub-markets and disregarding the degradation of Census representativeness over time. Chapters 5 and 6 explore methods for modelling sub-markets, whilst the following section discusses alternatives to Census data.

4.7 Other Sources of Demand Side Data

As the National Census of Population is only performed every 10 years, it is reasonable to expect degradation of this data over time. New circumstances can affect an area between successive Censuses, of which common examples are inward investment; employment changes; and improved communications.

It should be noticed that the empirical work presented in this thesis was based on Census data for four reasons: (1) physical ease of access; (2) used by others - geodemographics industry and UK house price indices; (3) universally comprehensible; (4) temporal correspondence with property database. The methods developed and the conclusions made are intended to be generic to other sources of demand side data - such as those supplied by other government statistics, market surveys and subjective opinions elicited from professional valuers (Almond, et al, 1997).

4.8 Conclusions

Literature suggests that location is the primary influence on residential property value (Mackmin, 1994). However, location in itself is not something that can be strictly defined as can number of bedrooms or the existence of a garage. Location is subjective, metamorphic and often unique. Professional appraisers must spend time getting accustomed to an unfamiliar location in order to appreciate and correctly interpret all aspects of a society (Mackmin, 1994). It is without doubt therefore that qualitative measure of location, neighbourhood and local economies form an integral part of any computer assisted residential appraisal model.

The average values of similar properties in a neighbourhood would seem to be a good place to start. Indeed, the empirical evidence presented in this chapter supports this approach. However, to gain a good measure of value ranges in an area

requires a good cross-section of representative data. Furthermore, the average values only really make sense if they are constructed for similar types of properties. This method then becomes similar to the DCC method that it is trying to support and thus suffers from data scarcity.

A description of a location's wealth, amenities, housing stock etc. could be a useful addition to an appraisal model. This level of information is available in the form of raw Census data and summarised Census data (geodemographic systems). The inclusion of Census data into a computer assisted appraisal model improved the accuracy of the model by an average of 2% using data at the District level and 7% using data at Enumeration District level.

By accessing more dynamic sources of data and including subjectively assigned variables from professional valuers, it is believed that appraisal models can be significantly enhanced.

5. STRATIFICATION USING CLUSTERING TECHNIQUES

This chapter builds on observations from previous work and investigates a method of stratifying the property market by grouping together geographical areas that share similar Census characteristics. Preceding an account of empirical work, the problem of modelling heterogeneous property markets is formally defined using predicate mathematics.

5.1 Introduction

A heterogeneous property market contains a range of different housing types across stratas of varied cultures, economies and environments. To assume that there is a single, smooth, and continuous, valuation function underpinning this complex interaction of demand and supply side variables in a heterogeneous market would be an oversimplification. In fact, this assumption is seldom made. Rather, mass appraisal researchers employ a number of methods to simplify this complexity, of which the following are the most common amongst published literature:

- manual selection of homogeneous areas;
- imputing rank by summarising categories of transactions;
- direct coupling of property transactions with other descriptive data;
- automated stratification into homogeneous subsets.

5.2 Manual selection of Homogeneous Areas

In this approach, data is carefully selected from a well-defined area believed to be homogeneous (homogeneity in this case is usually defined with respect to location and, therefore, encompasses environmental and local economic influences). This approach has been used to assess the usefulness of MRA, ANNs, Linear Programming and Expert Systems (Adair, et al, 1996; Evans, et al, 1992; Wiltshaw, 1991a; Gronow and Scott, 1985). Varying degrees of success have been achieved, and contributions to knowledge made with respect to coding and representation of

data (Borst, 1991; Gronow and Scott, 1985); selection and transformation of influential variables (Adair, et al, 1996; Greaves, 1984); and generally introducing data analysis tools to the valuation community.

Unfortunately, manual stratification is limited by the inherent complexity of the market. To properly construct a homogeneous space requires a complete understanding of the inter-relationships between the demand and supply side variables. Furthermore, it may be over zealous to create a model for each physical homogeneous region as this ignores other regions that share all the identified characteristics, which should, if appropriate, become integrated into a more complete model.

5.3 Ranking Location by Average Property Value

One method of representing location is to include in the appraisal model the average house prices for each location. Jenkins (1992) took this approach, where average values were fed into an expert system in order to provide a surrogate for location. This approach, as described in Chapter 4, was also taken for residential property data selected from the Cardiff area, using an ANN model. The results show a small increase in accuracy, most notable when average values were computed for house type within region. However, the relatively small increase in modelling accuracy suggests average value is not a sufficiently good surrogate. Yet, the approach is similar to that taken by many house buyers (and perhaps also estate agents) when assessing the relative 'value for money' of the subject property.

5.4 Coupling Property Data with Regional Statistics

To increase the modelling susceptibility of the data, it may be profitable to couple the property transaction data with other descriptive data sets. Data describing a region's wealth, employment, schools, air quality, weather, population etc. may prove useful as additional features to include in a model (see Mackmin, 1994 pp 5-34 for a more in-depth discussion). Sources of such data include: market research, regional employment statistics, census of population, regional crime statistics, school league tables, pollution studies, electoral roll, television regions, credit-related data and others. This approach was adopted in Chapter 4. Only 660 properties out of a total of over 1000 where used in the ED level analysis due to missing or incomplete values in the postcode attribute. Some improvement in accuracy was gained even at the

very highest level of abstraction. However, as had been anticipated, the greatest gains occurred at the ED level, where the mean absolute percentage error was reduced from 20% (original data set) to 13% (Original data and ED level Census statistics) for a randomly selected test set.

5.5 Automated Stratification into Homogeneous Subsets

Valuation literature presents a number of arguments supporting the existence of identifiable sub-markets within the residential property market. Strazheim (1973) suggests *"the urban housing market is, in fact, a set of compartmentalised and unique sub-markets delineated by housing type and location"*. Supporting research suggests that the heterogeneous market is best modelled by stratifying it into its underlying sub-markets. However, the composition of these sub-markets is neither singly nor well defined. The following is a summary of suggested stratification strategies:

Stratification by Location: The aim of stratification in this case, is to segment the heterogeneous property market into homogeneous sub-markets with respect to locational and environmental aspects. The underlying assumption is that the homogeneous sub-markets will contain less variance than the heterogeneous market and thus be more susceptible to accurate modelling (Adair and McGreal, 1995). This approach has been taken by a number of researchers (Sauter, 1985, Eckert, 1990, Adair and McGreal, 1995, Adair, et al, 1996) and in the main, the results have shown that the sub-markets are more susceptible to accurate modelling than the overall market. However, care must be taken in ensuring that boundaries are carefully selected and a significant sample size is maintained (Eckert, 1990).

Stratification by Property Type: Here, the property market is partitioned with respect to property characteristics. Properties are grouped according to size, age, type, number of storeys (Adair and McGreal, 1995) number of rooms, state of repair, number of bathrooms, garage (Have, et al, 1998).

Stratification by Buyer Behaviour: In the third case, described by Adair and McGreal (1995) as the 'behavioural' approach, segmentation is performed with respect of purchasers that are grouped according to *"identifiably different approaches to and valuations of various attributes"* (Adair and McGreal, 1995).

Stratification is generally performed manually, with attention given to a priori local knowledge. Adair and McGreal (1995) in their study of the Belfast residential property market, proposed that primary sub-markets should be formed for: Inner City; Middle City; Outer City; Greater East and South; Greater North; and Greater West. They also proposed that these sub-markets should then be further sub-divided into Terraced; Semi-detached; and Detached. However, the manual approach to stratification involves a high level of market understanding for the subject region, with each region having its own peculiarities.

Almy, et al, (1998) suggest that stratification of markets should be a preliminary step for all Computer Assisted Mass Appraisal (CAMA) systems. However, little attention has been paid to this type of automatic process in valuation literature.

Before considering approaches for automating the stratification process, it is useful to develop a visual interpretation of the role of sub-markets in a heterogeneous market. Figure 5.1 provides a purely abstract view of the interplay of sub-market functions in an heterogeneous market. Here a heterogeneous market is viewed as a conceptual mathematical space containing many functions accounting for the observed sub-market behaviour. The theoretical aim of any stratification process is then to segment this multifunctional space into smaller sub-regions containing a single value function.

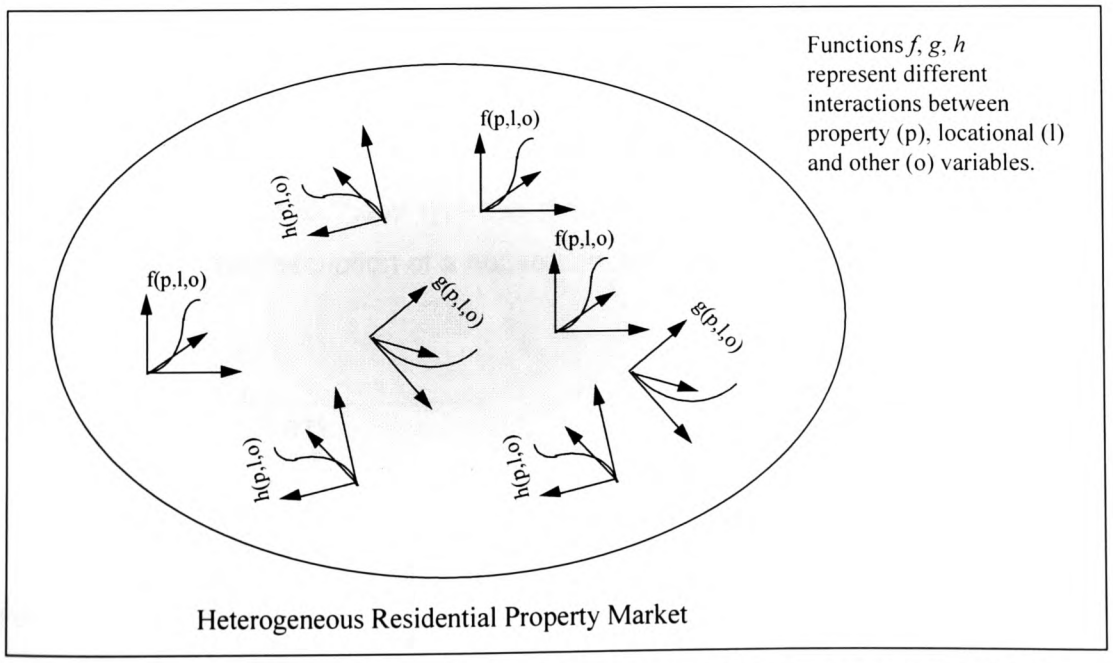


Figure 5.1 - Abstract Interpretation of Functions in a Heterogeneous Property Market Described in a Mathematical Conceptual Space.

For modelling purposes, it is also beneficial to develop a mathematical understanding of the underlying structure of the property markets. Although this structure is somewhat ambiguous, the theory is sufficiently well developed to build a model based on predicate logic. Section 5.6 defines a heterogeneous property market using the formal specification method Z (See Spivey, 1992).

5.6 Defining the Property Market Using Predicate Logic

In order to define an heterogeneous area in terms of homogeneous sub-regions and the houses within those sub-regions, it is useful to define 2 given types:

ATTRIBUTE: a set of attributes that can be used to describe a house;
PLACE: geographical identifier.

Attributes of a House: A house can be described in terms of a finite set of attributes. These attributes can be grouped into three main types: property (number of bedrooms, type, garage etc.); location (employment, schools etc.); and other (market trends etc.). This can be defined as:

PROPERTY, LOCATION, OTHER: ℙATTRIBUTE

PROPERTY ∩ LOCATION = ∅
PROPERTY ∩ OTHER = ∅
LOCATION ∩ OTHER = ∅

This predicate ensures that there is no overlap of attributes from one type to another.

The condition applied to these types is that there are no attributes shared by the derived types. The description of a house can then be formed in terms of its housing attributes:

HouseAttributes

property: ℙPROPERTY
location: ℙLOCATION
other: ℙOTHER

For example, ℙPROPERTY declares that the property attributes of a particular house is a subset of all possible property attributes relating to houses.

Obviously, a house possesses more than just its description - it also has a value. A basic valuation function can be defined that maps housing attributes to value:

BasicValuationFunction: HouseAttributes $\rightarrow \mathbb{Z}$

The basic valuation function can be read as taking a set of property, locational and other attributes relating to a house and returning the value of that house.

Theoretically, if all aspects of a house and their relative impact on value are known then this mapping should be true for all houses. However, in practice, the derivation of such a mapping is currently beyond the scope of the available modelling tools. Hence, a more realistic approach is to define a valuation function for each homogeneous sub-region, and group together those sub-regions that share the same valuation function. A sub-region can be defined in terms of its house types and its geographical location:

SUBREGION

theHouseTypes:	HouseAttributes
thePlace:	ℙPLACE

This allows the valuation function to be redefined as a higher order function, which when given a region will apply a function to the housing attributes and return the property value. This new valuation function is defined as:

ValuationFunction: SUBREGION $\rightarrow$ (HouseAttributes $\rightarrow \mathbb{Z}$)

The more detailed valuation function can be read as within a specified region, there is a function that given a full description of a house can return its market value.

This allows a particular house to be defined in terms of its attributes, region and value. Furthermore, the particular valuation function associated with the region when applied to the housing attributes returns the property value. This is defined as:

House

ItsHouseAttributes:	HouseAttributes
ItsValue:	$\mathbb{Z}$
itsRegion:	SUBREGION

ValuationFunction (itsRegion) ((itsHouseAttributes) = itsValue)

Defining Sub-Regions by Stratifying the Heterogeneous Space

The aim of stratification is to identify the sub-regions that share the same valuation function and group these into a single model. The definition of a sub-region therefore depends on the factors that define the homogeneity of an area of which the most fundamental are location and property type (Mackmin, 1994).

Stratification by Property Attributes

Here a sub-region is defined as a collection of houses sharing the same property attributes.

HomogeneousStrataByPropertyType

theSubRegions: ℙSUBREGION

$\forall S_1, S_2$: theSubRegions | S_1.thePlace ≠ S_2.thePlace ∧
S_1.theHouseTypes.Property = S_2.theHouseTypes.Property
$\forall h_1$: S_1.theHouseTypes, h_2: S_2.theHouseTypes •
ValuationFunction (S_1)(h_1) = ValuationFunction (S_2)(h_2)

This predicate proposes stratification by property type as being the means of achieving homogeneous sub-regions and reads as: for pairs of sub-regions with the same property types, identical homes will have the same value.

Stratification by Locational Attributes

Here a sub-region is defined as a collection of houses sharing the same locational attributes.

HomogeneousStrataByLocation

theSubRegions: ℙSUBREGION

$\forall S_1, S_2$: theSubRegions | S_1.thePlace ≠ S_2.thePlace ∧
S_1.theHouseTypes.Location = S_2.theHouseTypes.Location
$\forall h_1$: S_1.theHouseTypes, h_2: S_2.theHouseTypes •
ValuationFunction (S_1)(h_1) = ValuationFunction (S_2)(h_2)

This predicate proposes stratification by locational attributes as being the means of achieving homogeneous sub-regions and reads as: for pairs of sub-regions with the same locational characteristics, identical homes will have the same value.

Practical Interpretation

Unfortunately, due to the subjective and unique nature of property value, perfect stratification is not obtainable. However, sub-optimal stratification will undoubtedly go some way to improving the modelling capabilities. One approach considered worthy is the clustering of heterogeneous data into homogeneous subsets.

5.7 Stratification using Clustering Techniques

Geodemographic indicators could be employed to describe sub-regions in an heterogeneous residential area. Those sub-regions sharing the same characteristics could be grouped into a single model based on the assumption that similar areas have similar underlying value functions. Extending this reasoning, clusters found in Census data may correlate with homogeneous regions with respect to location, and clusters found in property data may describe homogeneous regions of properties. Early work by James (1994) concluded that an unsupervised neural network might be able to discern groupings within a parent data set that might represent homogeneous areas.

Using an Unsupervised Neural Network

An unsupervised network, such as the Kohonen network, organises itself in such a way as to represent classes within a data set. The 2-D Kohonen network allows classes to be visualised on a feature map, in which similar inputs are spatially clustered. Figure 5.2 shows a typical 2D Kohonen Self-Organising Map (SOM) along with an abridged algorithm (Note, the number of nodes are arbitrarily selected for example purposes).

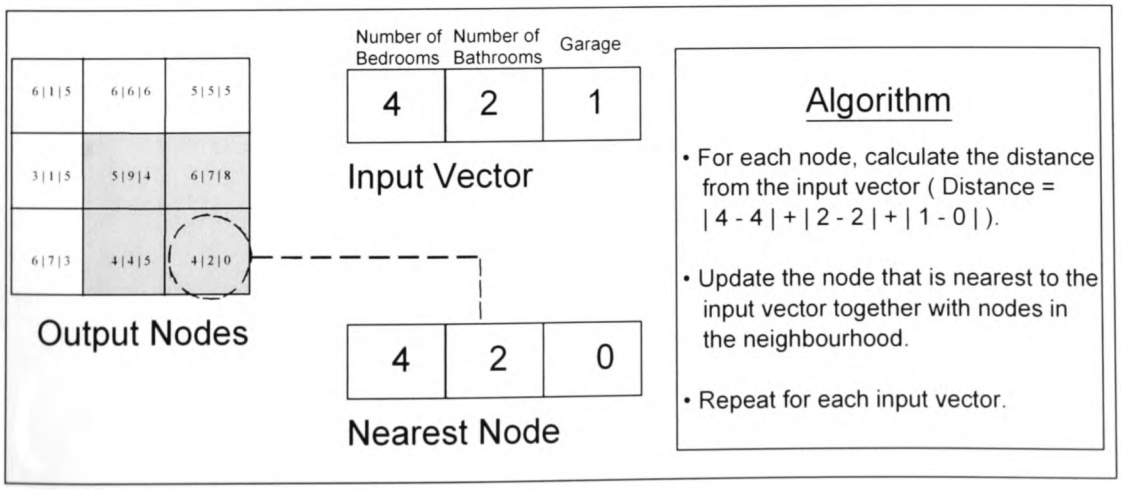


Figure 5.2 - A Kohonen Self Organising Feature Map.

Each output node on the Kohonen Feature Map contains a vector of length 'j', where 'j' is equal to the number of input attributes. Before training, the network is in an initialised state (i.e. the directions of the vectors in each node are random). Training involves passing an input vector into the network through the input nodes. Each node on the Kohonen Feature Map is then compared with the input vector, and the closest node is then changed to be more like the input vector. Neighbouring nodes also become more like the input vector. Iterating this process achieves clustering of similar input vectors in Euclidean space (Kohonen, 1984).

5.8 An Overview of the Methodology

The methodology uses a Kohonen Self-Organising Map (KSOM) to find clusters in the input vectors and then the data from each cluster is used to train a separate MLP network. The advantage of using the KSOM for this application is that it can identify clusters within the parent data-set that are difficult to identify using simple sort procedures.

Figure 5.3 gives an overview of the method. A data-set containing the required elements of the vector v is passed through the KSOM during the training stage and allowed to develop into clusters. After training, the clusters are inspected and the primary clustered features used to describe the sub-sets. Finally, vectors mapped to a cluster are separated from the parent data-set and used to train an independent MLP network.

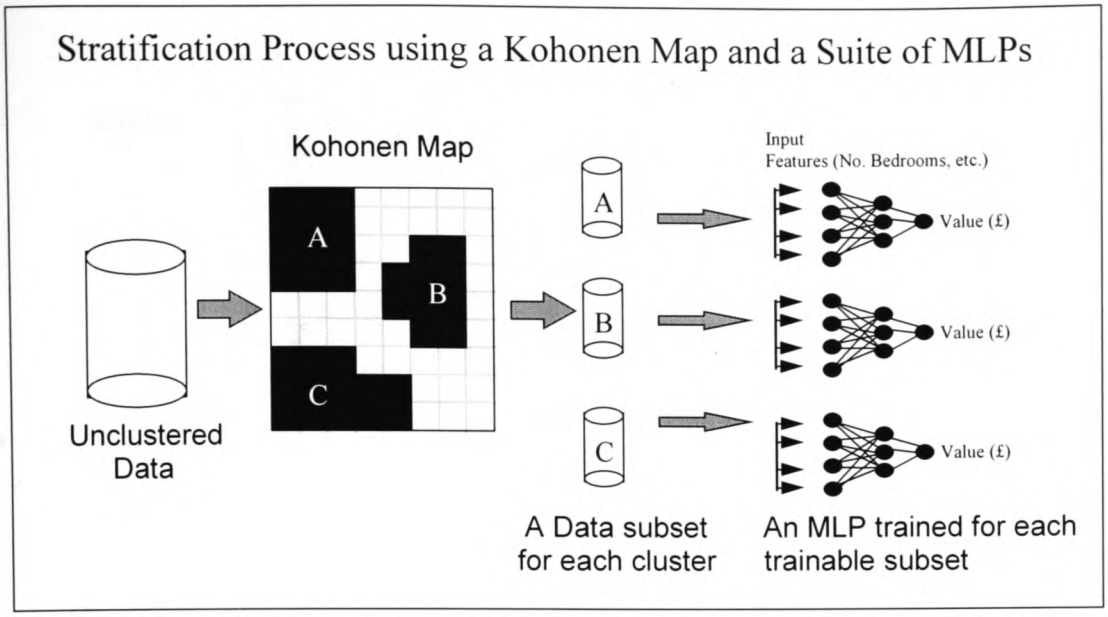


Figure 5.3 - An Overview of the Methodology. During Training, the whole historical data-set is separated - using a Kohonen Self Organising Map - into subsets that are subsequently used to train a series of multi-layered perceptron networks. During operation, the Kohonen Feature Map is used to determine which network to use to provide an estimate of value.

There are however two fundamental problems that need to be resolved before this method can be of any use. First, the problem of separating adjacent clusters, and second, the desire to proceed to the final stage only for clusters that describe 'good' training sets.

The first problem has been recognised in other studies (James, 1994) and some guidelines have been provided. In essence, the problem lies in the attribution of boundary nodes to a specific cluster. Figure 5.4 provides an example of a KSOM output with adjacent clusters. There appear to be four classes within the data-set, but there are regions of uncertainty relating to the boundaries of each cluster. (The Digits in each node show the number of vectors mapped to that node.)

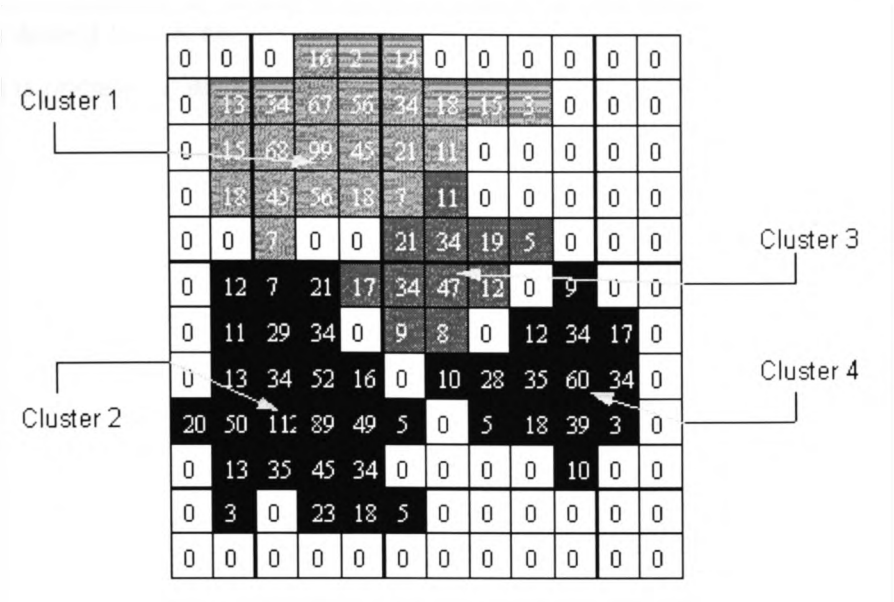


Figure 5.4 - An Example of a Trained Kohonen Self Organising Feature Map

To overcome this problem a simple method of identifying class boundaries or discriminants can be used, which relies on the fact that a KSOM clusters primarily on binary features. For example, if the type of house is represented using binary inputs, the KSOM will tend to cluster records according to this attribute. Boundaries between adjacent clusters on a 2D map can then be found by inspecting the records mapped to each node and grouping together nodes that contain the same classification values. However, this level of clustering can be achieved using a multi-level sort procedure. In essence, the binary representation of the data will dictate the make-up of the resulting clusters and more importantly the homogeneity of the data sets.

If the data are represented using continuous inputs, the clusters formed by the KSOM would provide more generalised classes which would be difficult to achieve using a sort procedure. However, the inspection method would no longer identify class boundaries, as the similarities between records would not always be apparent. Clearly, before meaningful training data sets can be formed, the problem of discerning effective class boundaries in a Kohonen feature map must be addressed. Ideally, the network adaption rule should cluster similar inputs and clearly distance individual clusters. Zurada, 1992 explains: *"One possible network adaption rule is: A pattern added to the cluster has to be closer to the centre of the cluster than to the centre of any other cluster"*. Using this rule, each node can be examined and the

distance from the surrounding centroids[3] can be calculated. The subject node can then be added to the nearest cluster. Figure 5.5 illustrates a hypothetical situation where it is unclear where to draw the boundaries around clusters on a Kohonen map.

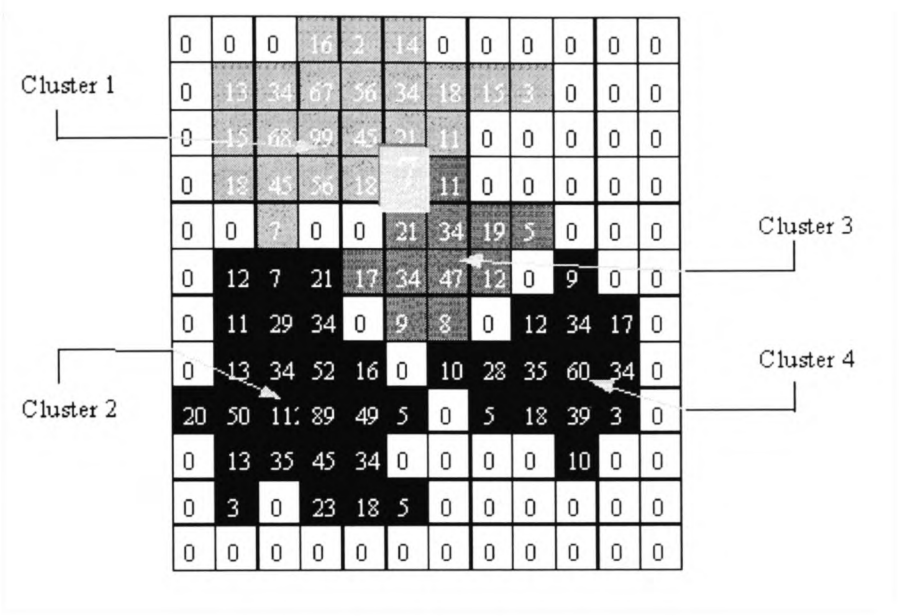


Figure 5.5 - An Example KSOM.

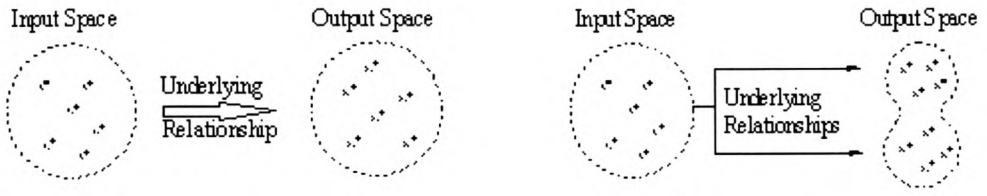
By simply calculating the Euclidean distance of the subject node from the two centroids, the subject node can be assigned to the cluster that is closest - for example Cluster 1. However, in this application, which aims to generate useful training data sets, the formation of a class boundary for Cluster 1 (including the subject node) may dramatically increase the variance of the training data. This increase will in turn reduce the potential accuracy of the back propagation model. In the example, it may have been better to exclude the subject node from either of the clusters, and deem the vectors mapped to the subject node as either being outliers or a separate cluster.

In addition to identifying boundaries around input clusters, it is also important in this application for input clusters to be matched by corresponding output clusters. In terms of residential property appraisal, if, for example, the Kohonen map has

[3] A centroid is taken to be a node that has the largest number of input vectors mapped to it.

clustered residential properties from two different locational areas, it is reasonable to expect similar types of houses from each area to have a similar property value.

Figure 5.6a shows a cluster of similar input vectors. When the corresponding data in output space is examined all the examples describe similar output values. For example, if the input cluster describes houses that have three bedrooms; semi-detached; less than 2 years old - then the cluster in output space says they all have similar property values. Conversely, Figure 5.6b shows a situation where the data can only be modelled using two or more functions.



(a) *(b)*

Figure 5.6 - Example Cluster Mappings from Input to Output Space

(a) An Example of a Good Input Cluster. A One-to-One relationship can be established and hence Input Space is homogeneous

(b) An Example of a Bad Input Cluster. Two or more identical vectors in the Input Space map to different vectors in the Output Space and hence, the Input Space is not homogeneous.

In order to determine whether the whole of the input and output space is suitable for being modelled with a single MLP - or whether data stratification needs to be performed - the algorithm outlined in Figure 5.7 can be applied:

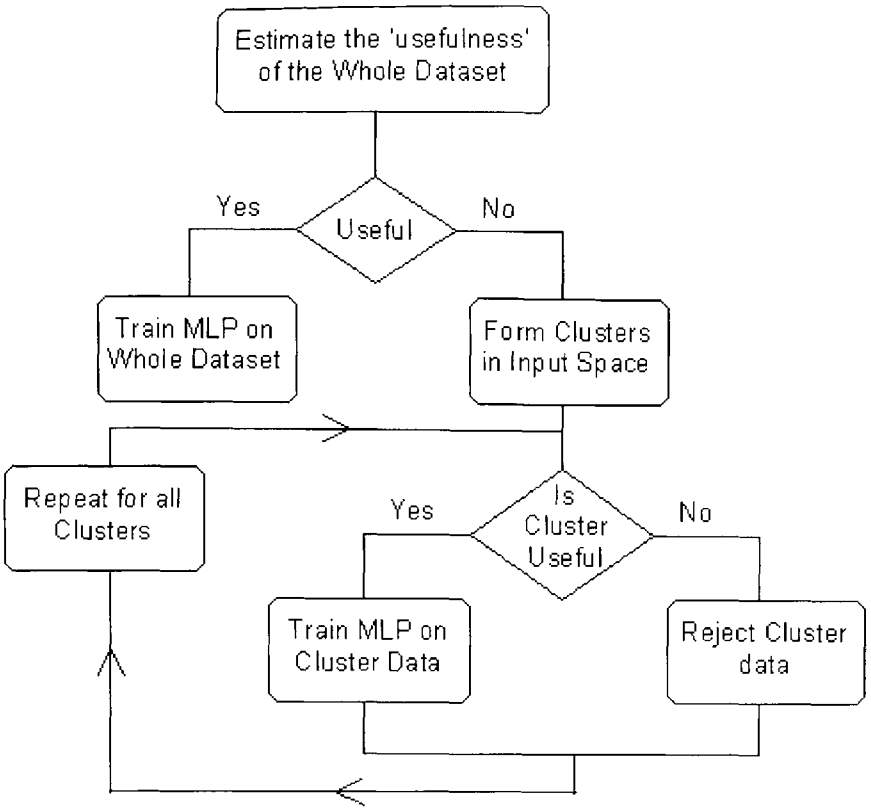


Figure 5.7 - Algorithm for Implementing Stratification of Training Data-sets.

The problem now is to estimate the 'usefulness' of a given cluster. There are a number of options available of which the following are the most useful:

- MLP Model (Chen, et al,1997)
- Class Entropy (Quinlan, 1986)
- R^2 (Almy, at al, 1998)
- Gamma Test (Stefannson, at al, 1997)

Obviously, the best estimate of how well an MLP can learn the underlying function in a training set is to actually train such an MLP using optimal parameters. Chen, et al, (1997) proposed this method in his stratification technique. However, the drawback with this technique lies in the setting of optimal parameters for the MLP. Any automatic implementation of the algorithm shown in Figure 5.7 would require strict rules governing the selection of transfer function, hidden layers / nodes and training

time. Although some formulas have been suggested to aid with such selection, this still remains in the main an empirical approach.

For classification problems, Class Entropy can be used to decide if input clusters are homogeneous with respect to output clusters. For example, Quinlan's C4.5 and C5.0 (Quinlan, 1993) uses Class Entropy to segment the input space until each segment points toward a single class in output space. However, this approach is not applicable for regression problems such as the derivation of residential property value.

A crude estimate of the susceptibility of a data-set for function induction can be achieved by executing a multiple regression analysis on the data and use the R^2 value to discern trainable clusters. This technique is useful for data where the function is known to be linear. However, this is not known to be true for residential property data, and in fact sufficient evidence exists to suggest the converse (Lam, 1996; Bruce and Sundell, 1977).

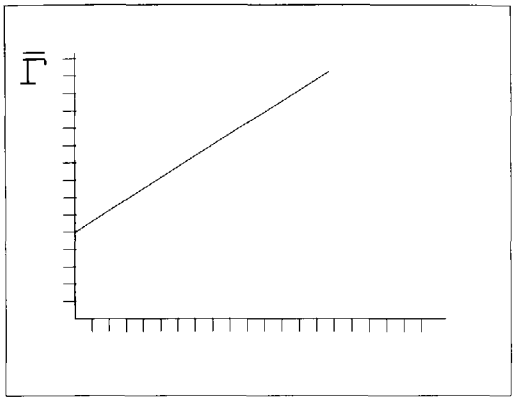
The Gamma Test

The Gamma test attempts to estimate the best mean square error that can be achieved by any smooth modelling technique using the data. If y is the output of a function then the Gamma test estimates the variance of the part of y that cannot be accounted for by a smooth (differentiable) functional transformation. Thus if $y = f(x) + r$, where the function f is unknown and r is statistical noise, the Gamma test estimates Var(r).

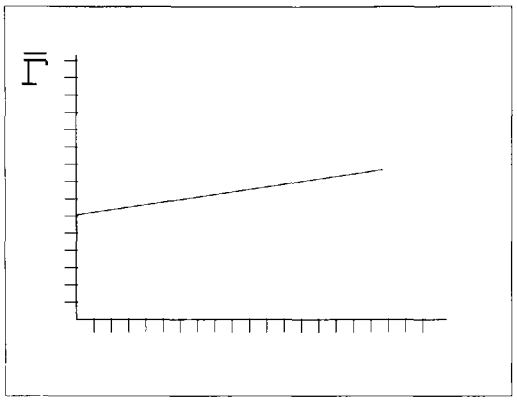
Var(r) provides a lower bound for the mean squared error of the output y, beyond which additional training is of no significant use. Therefore, knowing Var(r) for a data set allows prediction beforehand of what the MSE of the best possible neural network trained on that data would be. The Gamma test provides a method of determining the quality of the data stratification - a good stratification technique will result in a low value of Var(r) for each subset.

Interpreting the output from the Gamma test requires considerable care and attention. The least squares regression line provides two pieces of information. First, the intercept on the Gamma axis is an estimate of the best MSE achievable by any

smooth modelling technique. Second, the gradient gives an indication of the complexity of the underlying smooth function running through the data[4]. The Gamma test may estimate a very low MSE but unfortunately show a high level of complexity that could cause problems for a standard MLP network. It is easier to see this situation when the output from the Gamma test is presented graphically. A hypothetical example with high noise content and high complexity is shown in Figure 5.8(a); high noise and low complexity in Figure 5.8(b); low noise and high complexity in Figure 5.8 (c) and finally low noise and low complexity (the desired outcome) in Figure 5.8 (d).

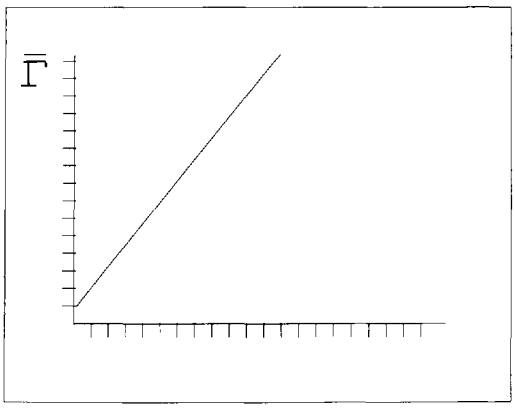


(a)

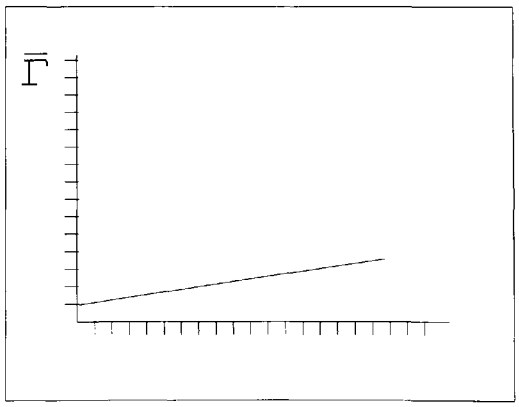


(b)

Figure 5.8 - Interpreting the Output from the Gamma Test
(a) High noise (large Gamma value) and high complexity (steep gradient)
(b) High noise (large Gamma value) and low complexity (flat gradient)



(c)



(d)

Figure 5.8 - Interpreting the Output from the Gamma Test
(c) Low noise (small Gamma value) and high complexity (steep gradient)
(d) Low noise (large Gamma value) and low complexity (steep gradient)

[4] This interpretation is based on empirical evidence and discussions with the research team who pioneered the Gamma test.

In summary, for this methodology to be successful, the following is required:

- class boundaries must be identified around clusters formed by the Kohonen feature map over the input space that exclude outliers and nodes from neighbouring clusters, and;
- only 'good' clusters (see Figure 5.6a) should go on to form training data sets for subsequent back propagation models.

A Detailed Look at the Methodology

The Gamma test can be used at a number of abstraction levels within the Kohonen stratification method:

- Cluster level
- Node Level
- Record Level

Data stratification is achieved at cluster level or at node level, depending on the ease at which cluster boundaries can be determined. The record level gives an indication of outliers.

Cluster Level Analysis: This can be achieved thus:

```
Identify Cluster boundaries in Kohonen map
For every cluster
        Place records mapped to cluster into a file
        Apply Gamma test to data in the file
        If Var(r) <= some Threshold then
                Use data file as the training set for MLP
        else
                Process at Node Level
```

This level of abstraction is the least computationally intensive as it only requires one pass of the Gamma test for each cluster. The disadvantage with this method is that it is often difficult to identify boundaries manually. In this case the Gamma test should be applied at the Node level.

Node Level Analysis: At this abstraction level, the methodology attempts to identify useful clusters by selecting a centroid and adding neighbouring nodes - where the

addition of a node increases the variance significantly, it is subsequently removed. This process iterates until the cluster size is maximised within a specified variance threshold. This is achieved thus:

```
Number_of_clusters:=0
While there are nodes to cluster

        number_of_clusters : = number_of_clusters + 1
        Select the unclustered node with the largest record count
        Apply Gamma test to estimate the variance for the data in the selected node

        If Var(r) <= Threshold then Nodes_of_interest:=None
       (Cluster includes only the data from selected node)

            For each unclustered node immediately surrounding selected node
                Add data from unclustered node to the cluster
                Run gamma test on cluster
                If Var(r) <= Threshold then
                        Add unclustered node number to nodes_of_interest
                Else
                        Remove data from the unclustered node from the cluster
                While nodes_of_interest <> None
                        Select c_node from nodes_of_interest
                        Remove c_node from nodes_of_interest
                        For each unclusterd node immediately surrounding c_node
                                Add data from unclustered node to the cluster
                                run gamma test on cluster
                                If Var(r) <= Threshold then
                                        Add unclusterd node  to nodes_of_interest
                                Else
                                        Remove data from the node from cluster

                                Record the boundaries of this cluster

        else Process at Record Level as node has too high variance to train an ANN
```

This algorithm identifies useful clusters on a 2D Kohonen map. The boundary detection algorithm for a 1D Kohonen map is very similar except neighbouring nodes are selected progressively further away from the left and the right of the centroid node.

This level of analysis is more computationally intensive than the cluster level analysis, as it require $m^*(n_i)$ passes of the Gamma test, where 'i' is the number of nodes investigated for cluster 'n' for a Kohonen map containing 'm' clusters.

If when using either the cluster level analysis or the node level analysis, useful clusters have been identified, it is then possible to train an independent MLP on each subset. The Kohonen map is then used to select the appropriate MLP on which to predict the value of a previously unseen example.

However, if both methods have still resulted in poor training sets (useless clusters) then the analysis is taken to the most detailed abstraction level, that is the record level.

Record Level Analysis: The record level analysis is the most computationally intensive. The purpose of this level of the methodology is to identify data subsets from examples that have mapped to the same node on the Kohonen map. This facilitates extraction of outliers from a data set as well as giving some indication as to the examples that require additional features.

The algorithm developed for this level of analysis is very similar to that shown for the node level analysis. However now, it is sets of records that are iteratively analysed using the Gamma test. This is achieved thus:

```
For each node in the Kohonen SOM
        Apply Gamma test to estimate the variance for the data in node
        If Var(r) > Threshold then
                For each record at node
                        Remove record from data set
                        Apply Gamma test to estimate the variance for the data in node
                        If New Var(r) < Previous Var(r) then
                                Mark record as outlier
                        else
                                Add record back into data set
        Else Proceed at Node Level
```

This level of analysis will identify the need for additional features and highlight records that may be classed as outliers.

5.9 Empirical Evidence

The empirical work investigated two of the stratification issues mentioned earlier: by property-type; and, by location.

Stratification by Property-Type

A 10 by 10 Kohonen SOM was used to find the groupings in the historical dataset containing 990 records. Value, the output attribute, was omitted from the data set used to train the Kohonen SOM. Using a combination of boundary detection methods, the data were found to contain eight groups. The records from each group

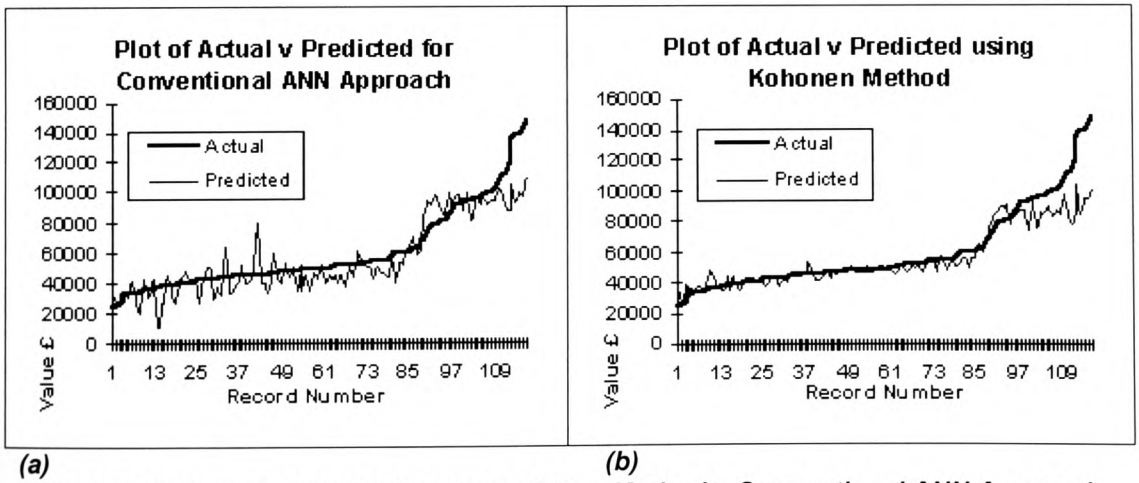
were examined and common attributes within the group removed. Obviously, as the dataset is partitioned into classes, the classes contain only a portion of the original 990 records. However, this is accompanied by a decrease in the number of attributes - as constant columns are removed.

In order to provide a benchmark for analysing the methodology, a single MLP model and a single MRA model were constructed using the whole data-set. After training the ability of the models to appraise residential properties with known values was tested. Table 5.1 illustrates the results achieved using the three methodologies on a test set of 117 properties.

Table 5.1 - Results achieved for the test set.

	Conventional MLP Method	MRA	Stratification Model
Mean absolute % error	18%	24%	8%
% of Records with an error > 10%	74%	79%	22%
Minimum absolute % error	0%	0%	0%
Maximum absolute % error	310%	220%	49%

Figure 5.9b shows the improvement in accuracy using the new method over the conventional Artificial Neural Network (ANN) approach (the data used for Figure 5.9a and Figure 5.9b have been sorted in ascending actual property value).



(a) *(b)*

Figure 5.9 - Improvement in Accuracy of New Method v Conventional ANN Approach

(a) A graph of actual and predicted value gained using a conventional neural network approach.

(b) A graph of actual and predicted value gained using the clustering approach.

It is evident, from the results obtained, that the methodology compares very favourably with the more conventional neural network approach and the multiple

regression approach. An average increase in prediction accuracy of 10% was achieved using the new method over the conventional approach. This implies that the original data-set either contained more than one underlying function (James, 1994) or the function was too elaborate to be modelled using a single MLP network. Moreover the Kohonen Feature Map can discern different classes within the data, which when independently modelled yield a greater predictive accuracy than those computed for the original data-set (James, 1994). In this way, the system may transcend the limitations implied by traditional locational versus sectoral searches. Sufficiently refined clusters may represent specific property types in specific sub-markets.

However, further analysis suggested that clusters formed by the Kohonen Self Organising Map did not always lead to dramatic increases in prediction accuracy. This was to be expected, as no real indicators of location were included in the analysis. To address this problem, 'demand-side data' as described in Chapter 4 was considered.

Training the Kohonen on Census aggregates

Training the Kohonen map on the Census data was relatively straight forward, however, the process had to be embedded within some control software in order for the post clustering fitness estimation to be made. Figure 5.10 presents a framework for including Census data in the stratification model.

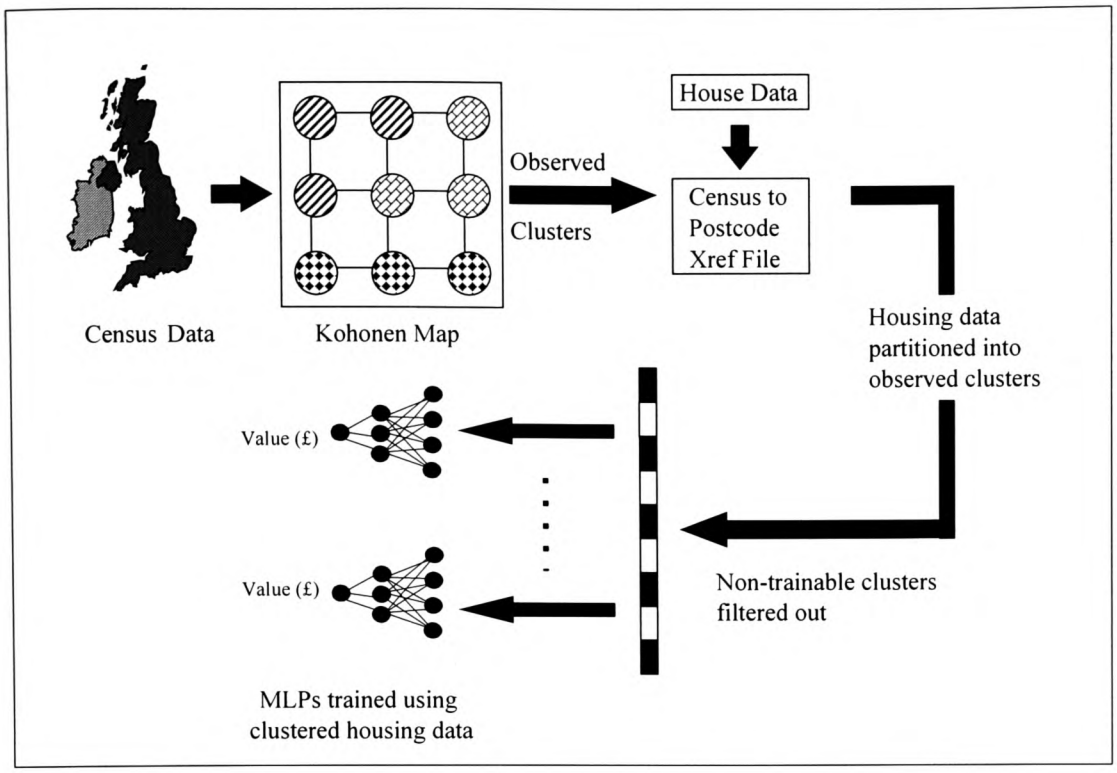


Figure 5.10 - A Framework for Including Census Data in the Stratification Model.

In order to test the effectiveness of the method for determining useful residential property sub-markets, two testing procedures were developed:

- sub-model v single control model
- sub-model performance on unseen 'similar area' data

Sub-Model v Single Control Model Results

Sub-models were constructed based on individual geodemographic factors, for example housing stock in observed region (collection of postal code regions) and employment statistics. The value of test properties were estimated using the single MLP model trained on all of the available data and also using the appropriate sub-models.

The results are presented in Table 5.2 through to Table 5.8. The following metrics have been used to compare the predictions made by the Single ANN model and the Kohonen stratified models

R^2 - This measure is used to estimate the degree of correlation between the model prediction and the actual property values. The following equation was used to compute R^2.

$$R^2 = \frac{\sum_{i=1}^{n}(P_i - \overline{A})^2}{\sum_{i=1}^{n}(A_i - \overline{A})^2}$$

SSq - This measure gives an estimation of the errors for the particular model. A lower value in this measure equates to a higher predictive accuracy. The following equation was used to compute SSq.

$$SSq = \frac{\sum_{i=1}^{n}(A_i - P_i)^2}{n}$$

COV - Covariance is an average of the product of deviations from one distribution to another. This measure also therefore provides an estimate of how closely the predicted values match the actual values. Again a lower value in this measure equates to higher level of accuracy. The standard formula to compute COV is:

$$COV = \frac{1}{n}\sum_{j=1}^{n}(P_j - \overline{P})(A_j - \overline{A})$$

MSE - Absolute percentage mean square error provides an estimate of accuracy that can be compared with the requirement of professional valuations to be within a 15% bounding of the market value. This measure is computed thus:

$$MSE = \frac{\sum_{j=1}^{n}\frac{100 * Abs(P_j - A_j)}{A_j}}{n}$$

Tables 5.2 through to 5.8 show the results of clusters indentified using the algorthms previously described with the Gamma test used to estimate the suitability of the data

as a training set for an MLP network (For a list of the Census attributes used see Appendix 1.2). The results are presented as sub-model value then single model value for each metric. For example an entry in the MSE column of 16.4 [20.0] represents a MSE value of 16.4 for the sub-model and 20.0 for the single model. Where the sub-model accuracy is greater than the single model the entry is highlighted. Graphical representations of these results can be found in Appendix 2.

Table 5.2 - Results for House Type Analysis

	MSE	R^2	SSq	COV	>15%
HT1	16.4 [20.0]	0.84 [0.68]	80 [117]	197 [202]	46 [46]
HT2	15.0 [43.1]	0.76 [0.72]	91 [689]	261 [-44]	46 [50]
HT3	14.73 [15.7]	0.95 [0.67]	102 [189]	583 [742]	41 [45]
HT4	13.9 [21.22]	0.99 [0.95]	48 [104]	266 [225]	38 [44]
HT5	20.2 [22.6]	0.28 [0.59]	380 [390]	174 [225]	45 [45]

The Kohonen Map was trained using Census variables describing the number of terraced, detached, flats, bedsits and semi-detached properties in the surrounding neighbourhood.

Table 5.3 - Results for Employment Analysis

	MSE	R^2	SSq	COV	>15%
EMP1	19.2 [19.8]	0.77 [0.56]	148 [170]	554 [468]	52 [53]
EMP3	15.7 [18.3]	0.99 [0.88]	82 [145]	344 [336]	44 [45]
EMP4	18.8 [20.6]	0.60 [0.48]	167 [184]	187 [230]	46 [52]
EMP5	16.7 [19.6]	0.92 [0.67]	98 [120]	152 [200]	43 [46]

The Kohonen Map was trained using Census variables describing the percentage of people in full-time and part-time employment and also those on government schemes and unemployed in the surrounding neighbourhood.

Table 5.4 - Results for Tenure Analysis

	MSE	R^2	SSq	COV	>15%
TEN1	16.1 [18.2]	0.87 [0.54]	137 [134]	152 [200]	48 [50]
TEN3	13.3 [23.0]	0.99 [0.75]	83 [128]	727 [611]	36 [49]
TEN5	19.8 [20.4]	0.82 [0.56]	131 [131]	169 [252]	46 [52]

The Kohonen Map was trained using Census variables describing the percentage of homes that were owner outright, owner buying, rented from county councils etc. in the surrounding neighbourhood.

Table 5.5 - Results for Car Availability Analysis

	MSE	R^2	SSq	COV	>15%
CAR1	13.3 [18.5]	0.90 [0.66]	620 [844]	222 [179]	34 [49]
CAR4	15.1 [17.7]	0.99 [0.52]	93 [119]	175 [256]	38 [45]

The Kohonen Map was trained using Census variables describing the percentage of families without a car, owning 1 car, owning 2 cars etc. in the surrounding neighbourhood.

Table 5.6 - Results for Ethnic Analysis

	MSE	R^2	SSq	COV	>15%
ETH1	13.3 [20.2]	0.72 [0.69]	61 [108]	208 [180]	32 [58]
ETH6	18.5 [20.1]	0.85 [0.55]	100 [113]	198 [239]	43 [46]

The Kohonen Map was trained using Census variables describing the percentage of families from various ethnic origins in the surrounding neighbourhood.

Table 5.7 - Results for Socio-Economic Analysis

	MSE	R^2	SSq	COV	>15%
SEG1	19.3 [21.5]	0.64 [0.60]	108 [132]	224 [219]	48 [51]
SEG5	15.2 [16.3]	0.97 [0.50]	118 [344]	764 [452]	47 [47]

The Kohonen Map was trained using Census statistics describing the social-economic group of households in the surrounding neighbourhood.

Table 5.8 - Results for Education Analysis

	MSE	R^2	SSq	COV	>15%
EDU1	18.2 [19.7]	0.96 [0.43]	119 [131]	200 [291]	45 [48]
EDU4	17.3 [19.3]	0.78 [0.58]	142 [158]	446 [377]	49 [52]
EDU5	16.4 [19.4]	0.70 [0.60]	75 [118]	191 [174]	40 [48]

The Kohonen Map was trained using Census statistics describing the percentage of households where members hold various qualifications such as degree, diplomas etc. in the surrounding neighbourhood.

Discussion of Results

The results generally show an average improvement in value estimation by the sub-model over the control model. Table 5.9 shows the average within-category errors for the sub-models:

Table 5.9 - Average Errors for Within-Category Sub-Model Estimates

Sub-Model Category	Single-Model Error	Sub-Model Error	Improvement
House-Type	19.36%	16.12%	3.24%
Employment	19.60%	17.69%	1.91%
Tenure	20.58%	16.44%	4.14%
Car-Availability	18.12%	14.29%	3.83%
Ethnic	20.15%	15.90%	4.25%
Socio-economic	19.10%	17.29%	1.81%
Education	19.53%	17.98%	1.55%
Average	19.49%	16.53%	2.96%

Table 5.10 shows the percentage of estimates for each category that has an error more than 15%.

Table 5.10 - Average Percentage of Estimates with More than 15 % Error

Sub-Model Category	Single-Model > 15%	Sub-Model > 15%	Improvement
House-Type	51%	45%	6%
Employment	49%	46%	3%
Tenure	50%	43%	7%
Car-Availability	47%	36%	11%
Ethnic	52%	38%	14%
Socio-economic	49%	48%	1%
Education	49%	45%	4%
Average	50%	43%	7%

Sub-Model Performance on Unseen 'Similar Area' Data

The second testing procedure determines whether a sub-model trained using data selected from one geographical area could effectively be used to predict the values of properties from a different geographical area that shares similar Census characteristics (See Appendix 1.2 for a list of Census attributes used).

Table 5.11 - Sample Results for Two-County Analysis

Data Sample	Mean Absolute % Error	Improvement
County_B Whole Data set	18%	-
County_A Model_A	18%	0%
County_A Model_B	17%	1%
County_A Model_C	15%	3%
County_A Model_D	15%	3%
County_A Model_E	18%	0%

The sub-models compared with the single model predictions achieved an accuracy increase of between 1% and 14%.

MLP models were created using property data selected from one county in South Wales (County_A) and were used to predict the values of residential properties in a neighbouring county (County_B). The results (see Table 5.11) show that the County_A models at least match, and in some cases outperform, the predictions made by the single MLP model trained on County_B properties. Clearly, the ability of these models to predict the values of residential properties outside the geographical area from which the training data was selected has been demonstrated.

5.10 Conclusions

From the research presented in this Chapter a number of conclusions can be made:

- A set of models, each dedicated to a certain narrow domain, can significantly outperform predictions made by a single more general model trained on all of the available training data.
- Models created from the stratification technique can be used to predict property values in other areas that have similar Census characteristics.

Based upon these results consideration has been made to the development of hybrid systems, where different technologies can be brought together into one system to compliment each other.

It is evident, from the results obtained, that the methodology compares very favourably with the more conventional neural network approach. An average increase in prediction accuracy of 10% was achieved using the new method over the conventional approach. This implies that the original data set either contained more than one underlying function (pattern) (James, 1994) or the function was too elaborate to be modelled using a single back propagation network. Moreover, the Kohonen network can discern different classes within the data, which when independently modelled yield a greater predictive accuracy than those computed for the original data set (James, 1994). In addition to this, analysis suggests the Kohonen step can be applied to subsets of the data (for example Class A) to create even more accurate sub-models.

The technique employed to estimate the susceptibility of the data to be modelled was the Gamma test. Based on a nearest-neighbour approach, this method gives a measure of both noise (intercept) and complexity (gradient), assuming a single smooth continuous function underpins the data-set. A number of assumptions can therefore be made about a set of data given its Gamma results. Firstly, a data-set with a high noise value may have insufficient examples, descriptive features or contain data mapped by multiple underlying functions. Assumptions can also be made based on the complexity value, a very complex underlying function may in fact be an aggregate of multiple functions which may be too complex for MLP or MRA to model.

Although these results are promising, the generation of suitable training sets relies heavily upon selection of useful neighbourhood characteristics. Poor selection leads to clusters forming which perform badly when analysed by the Gamma test. This method is therefore of most use when a priori knowledge is available to determine the neighbourhood descriptors to select.

6. STRATIFICATION USING SEARCH TECHNIQUES

Following the assumption that a heterogeneous property market can be modelled as a set of homogeneous sub-markets, this chapter illustrates a more dynamic approach to generating such sub-markets. The aim of this research was to reduce the dependency of the model on the original 'cluster' data and provide a method of navigating from a poor subset to one that could be used to define a training set for a regression based modelling tool such as ANNs.

6.1 Introduction

In the previous chapter, the heterogeneous property market was segmented using a clustering technique, resulting in an improvement in accuracy for those clusters with 'good' Gamma results. However, the disadvantage with the method related to the estimation of trainability post clustering. An example was given for a cluster - where for each property - the variable FrontDoorColour had the value White. This of course is unlikely to describe a useful sub-market of properties with respect to value. The alternative is to estimate trainability on-route. This requires the problem to be represented in such a way as to enable the clustering to be directed towards the 'best' formation. This type of approach is normally achieved using a state space representation with navigation through possible states guided by a search strategy.

6.2 State Space Representation

In classic state space representation, an initial state is created that contains all the facts of the unresolved problem. Successive states are generated using defined operators. This process is iterated for all successive states until termination occurs as a result of: exhaustive state generation; number of states exceeds some desired threshold; or, a goal state is reached. The developer of such a system must define: operators that can generate successive states; an upper limit on number of states; and, a goal state (solution). Navigation through the state space is controlled by a

search strategy, of which there are a number of alternatives with the most simple being:

- breadth first (states are generated and analysed horizontally);
- depth first (states are generated and analysed vertically);
- best first (states are generated from the current best state).

Figure 6.1 shows an example state space representation for an undefined problem, together with the route taken by breadth first and depth first search strategies.

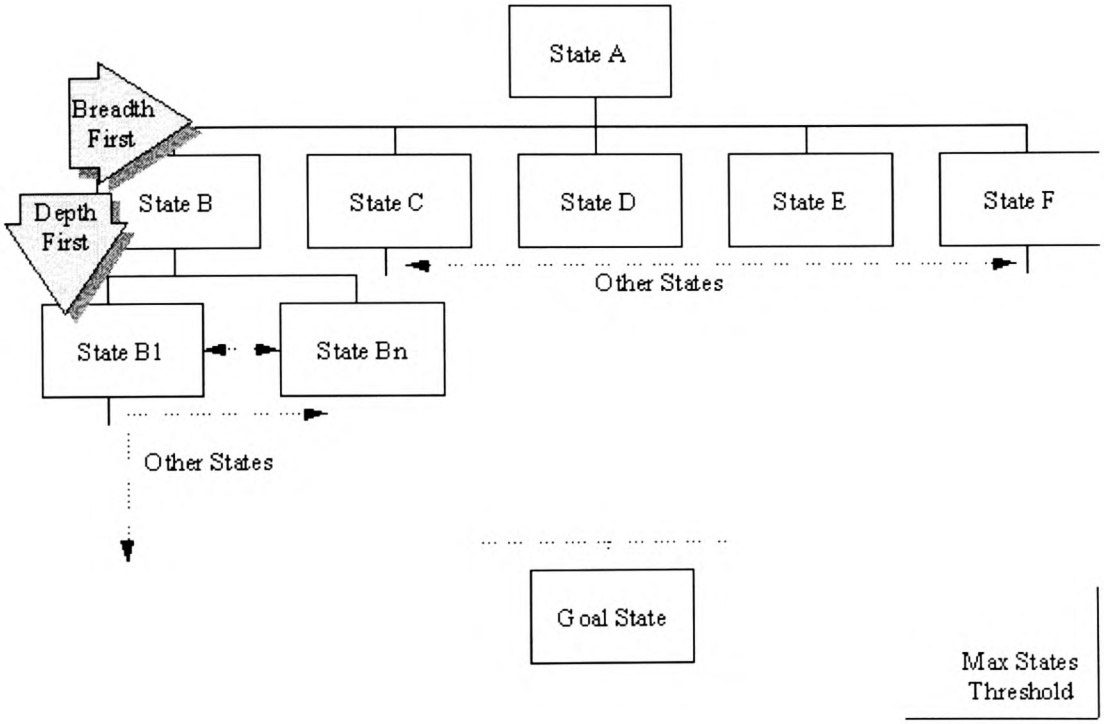


Figure 6.1 - Classic State Space Representation (Breadth First and Depth First Search Strategies)

Some problems are easily solved using either breadth first or depth first strategies. However, both methods 'blindly' search the state space in search for a goal state and can be extremely inefficient for large state space problems.

More efficient state space searching can be achieved by analysing only the best states and assuming these are closest to the goal-state. As before, operators are applied to generate new states. However, in this case each new state is compared to all known states using some state evaluation function (SEF). The state that shows

the most potential - based on its SEF value - becomes the current state on which the state generation operators are applied. Searching terminates, as with the previous strategies, when a goal state or other stopping condition is reached. However, termination can also occur when all additional states have worse SEF values than the current state. This type of termination is useful for problems where a goal state is difficult to define precisely.

Figure 6.2 provides an example of a state space representation using a 'best first' search strategy. The SEF is undefined to the reader but can be interpreted as the best state having the highest SEF value.

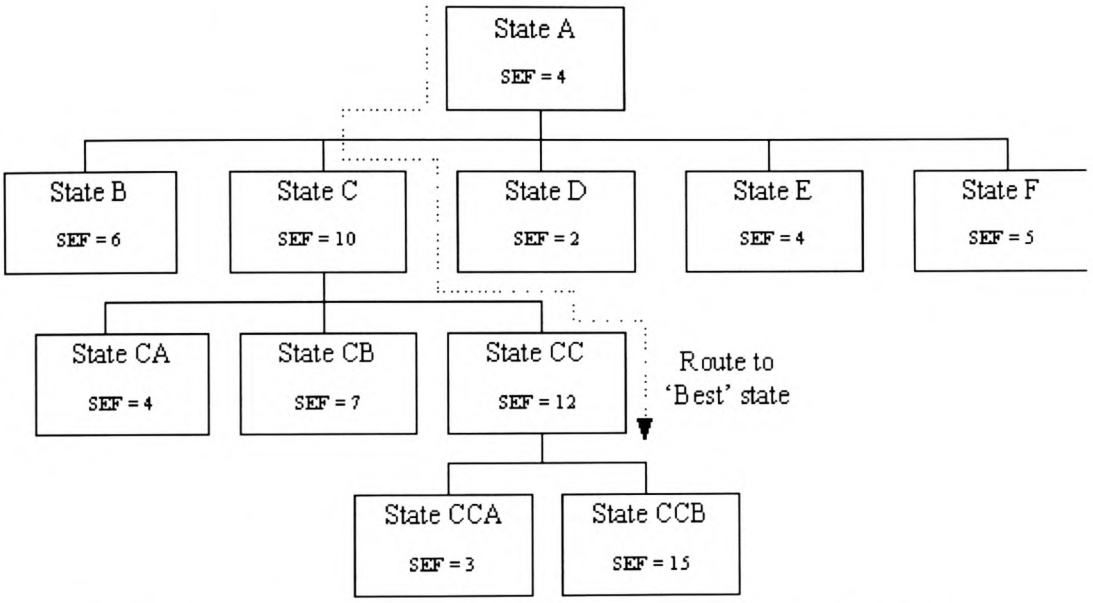


Figure 6.2 - Typical State Space Representation (Best First Search Strategy)

6.3 Tree Based Implementation of a State Space Problem

One method of implementing the abstract state space representation is to use a tree structure. For the stratification problem, the initial state, or the root, contains all the training data. Successive states are generated by partitioning the training examples into sub-sets and estimating the relative increase in modelling accuracy. This approach was used by Chen, et al, (1997). Stratification was achieved using linear discriminants and fitness estimated by training an MLP network on the current states and evaluating whether the MLP network was 'good' or 'bad'. The data used to train a 'bad' network was further analysed using the stratification technique.

This idea of divide-and-conquer is meritorious, especially for large-scale problems where data show signs of multiple functions. However, there are two main drawbacks with Chen's technique:

State Evaluation Function: The success of a MLP as a state evaluation function is relative to the network parameters chosen and the quality of the test set. The MLP could be replaced by an MRA model and the value of R^2 used to determine the trainability of the data. However, this forces the model to be linear in nature. A better approach would be to estimate the lowest mean-square error (MSE) present in the data regardless of idiosyncrasies associated with a particular modelling technique. This, as with the work In the previous chapter, is best achieved using the Gamma test (Stefánsson, et al, 1997).

Inefficiency of Tree Implementation: The computational requirement is large for the tree structure suggested as each new state is created by a well defined operator that normally acts in very small steps. Large state space representations can take many hours to arrive at a goal state.

For these reasons, it was decided that the problem was best approached - given its proven track record in large combinatorial problems - using a genetic algorithm. The fitness function selected was the Gamma test.

6.4 An Overview of Genetic Algorithms

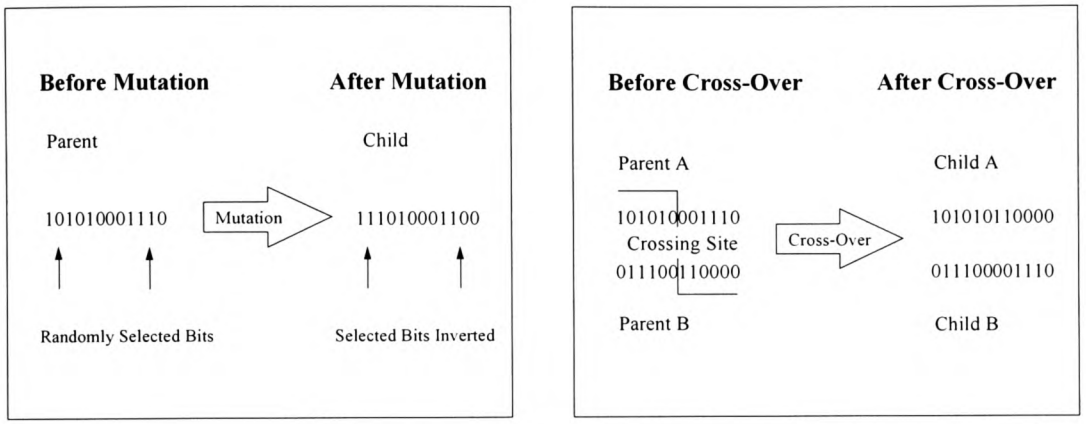
Genetic Algorithms (GAs) simulate the Darwinian theory of evolution (See Goldberg, 1989 for a good introduction to GAs). A typical GA operates at the level of genetic coding: the chromosomes or genotypes. Genotypes are usually simple bit strings of fixed length; the related 'adult' individuals (phenotypes) are obtained by decoding such bit strings using domain information. The basic steps of a typical genetic algorithm are as follows:

1. Randomly generate a population of individuals (bit strings).
2. Decode each individual and evaluate its fitness.
3. Generate a new population using cloning (survival of current individuals); cross-over (bit string reproduction); and mutation (random changing of bits in current individuals).

4. Repeat steps 2 and 3 until convergence (or another stopping condition reached).

Mutation and Cross-over Operators

The cross-over and mutation operators are fundamental to the development of a GA solution, from its random initial state to a near optimal mature state. In most GA applications, the encoding permits the use of standard mutation operators (random inversion of bits in a chromosome) and standard single or multiple cut crossover operators. Figure 6.3 illustrates the effect of applying a mutation operator (a) and a single cut cross-over operator (b).



(a) *(b)*

Figure 6.3 - Genetic Algorithm State Operators
(a) A Schematic of simple mutation showing the inversion of randomly selected bits.
(b) A Schematic of simple cross-over showing the partial exchange of information, using a crossing site chosen at random.

6.5 Stratifying Training Data Using a Genetic Algorithm

The GA approach benefits from having binary representations of the domain data. This allows the standard cross-over and mutation operators to be applied at bit level. The steps followed to formulate this stratification problem into one suitable for a GA to process were:

- recode the Census data describing each ED as a bit string;
- set up a look-up table to allow properties residing within an ED to be placed into a file suitable for analysis by the Gamma fitness test;
- Develop a GA application to generate stratas and test fitness;

Recoding the Census Data

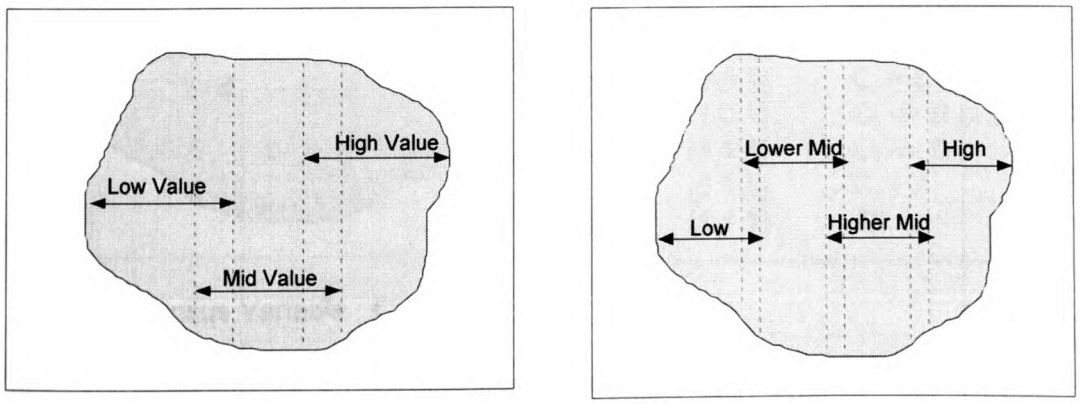
The first step taken to achieve a binary representation for the 1991 Census data was to form a discrete representation of the raw data by normalising between 0 and 100. For statistics relating to households, such as the number of terraced properties, this was achieved thus:

NormalisedValue = RawValue / NoOfHouseholds

For statistics relating to population, such as number of pensioners, the following equation was used:

NormailsedValue = RawValue / NoOfPersons

Each Census feature then represents percentage households or percentage persons in each ED. These discrete representations can be converted to binary representations by setting thresholds and using values of True (1) if the subject value is below the threshold and False (0) otherwise. For example, consider a Census variable C1 describing the percentage of terraced properties in an ED, with thresholds of C1<10% (low); 10%<= C1< 20% (mid); C1>=20% (high). An ED with 30% terraced properties could be described using 3 binary variables as: 0 (low) 0 (mid) 1 (high). These partitions can be improved by the addition of a fuzzy boundary, making the transition from one classification to the next less strict. This type of continuous valued thresholding is sometimes known as soft partitioning. Figure 6.4 (a) and (b) give examples of a Census variable split by 2 and 3 soft partitions respectively.



(a) *(b)*

Figure 6.4 - Continuous Valued Thresholding (Soft Partitioning)

(a) 2 Soft Partitions ***(b) 3 Soft Partitions***

Splitting each Census feature in this way allows a binary value to be used for each partition. For example, in the case of the 2 partitions an ED that has a value below the 'Low' threshold can be represented as [1,0,0] and mid and high value EDs can be represented as [0,1,0] and [0,0,1] respectively.

The thresholds could be set manually after an inspection of the data. However, given the size of the task, an automatic approach was favoured. The following equations were used to set the thresholds for all Census variables given any number of partitions:

$$\text{LowerBound}_i = (i-1)*(\beta-\alpha)+Min \quad (1)$$

$$\text{UpperBound}_i = LowerBound_i + Width \quad (2)$$

Where: i: Partition Number

$Width = 2\alpha + \beta$

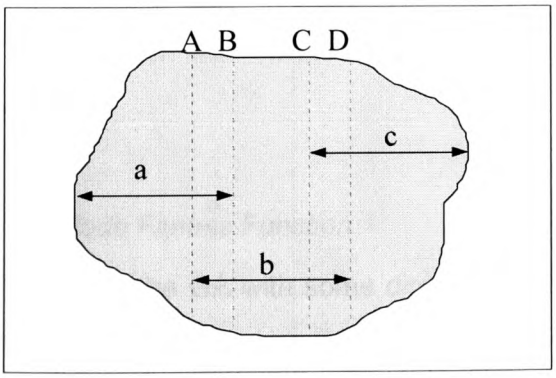
$\beta = \frac{Max - Min}{n}$

$\alpha = Overlap\ Ratio * \beta$

Max: Maximum Value for Feature

Min: Minimum Value for Feature

Figure 6.5 shows an example, where a Census variable has been divided using 2 partitions. All possible binary representations and their decoded meaning are given.



[a,b,c]	EDs Selected
[0,0,0]	All
[1,0,0]	C_1 <= B
[0,1,0]	A < C_1 <=D
[0,0,1]	C_1 > C
[1,0,1]	C_1 <= B or C_1 > C
[1,1,0]	C_1 <= D
[0,1,1]	C_1 > A
[1,1,1]	All

Example Census Variable : C_1

Figure 6.5 - Decoding the Binary Representations Found in an Example Using 2 Partitions.

Generating an Initial Population

Conventionally in GA systems, the initial population contains randomly generated bit strings of the required length, each of which can be decoded to represent a state in the problem space. This is not the case for this GA system, as each chromosome represents a set of Census characteristics of which only a subset of all possible chromosomes actually exist as EDs. Therefore, it is fair to assume that some members of a GA population may describe regions that do not exist in the selected ED data. This leads to a situation where the fitness function (Gamma test) cannot be applied as there are no training examples available.

To understand the difficulty of the task the GA has, it is necessary to have an appreciation of the size of the state space. Consider an ED described using 10 Census variables: If each variable is partitioned into 3 binary values, then in total there are 30 binary values describing an ED. This means that there are 2^{30} (1,073,741,824) possible states. To overcome this problem, three approaches were considered:

Random Initial Population

This 'suck-it-and-see' approach simply involves following the normal operating procedure for a GA system, using a randomly generated initial population. Fitness is evaluated using the Gamma test with '0' fitness returned for a non-existent ED.

This approach was however deemed to be inefficient, as it is essentially a random walk.

Dual Mode Fitness Function

To provide the GA with some direction, the following dual mode fitness function was defined:

if chromosome matched ED then Fitness = Gamma(ED)
else Fitness = Number of matching bits

This meant the chromosomes that were most similar to EDs were more likely to propagate into later generations.

Initial trials suggested that this approach provided the GA with reasonable navigation for a small number of Census variables (<=15).

Non-Random Initial Population

A third approach was investigated that forced the initial population to exist more closely to the existing EDs. In this approach, the initial population was formed by copying randomly selected EDs from the ED file. Before testing fitness and beginning the evolution process, each chromosome was mutated. This process worked reasonably well for large numbers of Census variables (>15). However, without a sufficiently large number of mutations, the final population remained relatively close to the initial one.

On the basis of some initial trials, it was decided to use the dual-mode fitness function whenever possible and resort to the non-random approach for analysis using more than 15 Census variables.

Decoding and State Evaluation

Chromosome decoding is required after each new generation is created in order to fix probabilities for survival (compete or partial) in later life. Decoding is relatively simple in this application with each (n-1) bits (where n is the number of partitions) representing a single Census variable. EDs that satisfy a chromosome description can be identified, and residential properties within the selected EDs can be grouped and placed into a separate data set. State Evaluation involves applying the Gamma test to the described data set and the appropriate metric(s) recorded. However, to speed up this process an initial pass of the Census data set is performed before the GA is run. The Census variables associated with each ED are converted into a binary representation using the described technique. Individual EDs can then be selected during the GA run, by matching binary strings ignoring any [1,1,1] and [0,0,0] sub-strings in the current generation. Figure 6.6 provides examples of this partial matching technique between chromosomes and EDs.

	A B C D E F G 12312312312312312 3	7 Census variables (A..F) divided into 3 groups using 2 soft-partitions
ED String Chromosome	100010001100010001100 100010001100010001100	Chromosome matches ED string exactly so ED is selected
ED String Chromosome	100010001100010001100 111010001100010000100	Chromosome matches ED for each Census variable except where chromosome string is [1,1,1] or [0,0,0]. Hence, ED is selected
ED String Chromosome	100010001100010001100 100010001101010000100	Chromosome does not match ED string so ED is not selected.

Figure 6.6 - Illustration of Partial Chromosome/ED Matching Used to Select EDs on the Basis the Current GA Solution

6.6 Empirical Evidence

Before commencing the GA analysis, the whole data set containing residential property transactions was analysed using the Gamma test. The results of this analysis are shown in Table 6.1.

Table 6.1 - Gamma Analysis of Whole Data set

Gamma Intercept	Gamma Gradient	Mean Absolute Percentage Error
0.0870	0.0284	20.57%

Although the interpretation of the Gamma test is non-trivial, it is reasonable to conclude from the relatively high intercept value that the mapping from input space to output space contains high noise levels. The noise can be attributed to: insufficient explanatory variables; multiple functionality; or, a combination of both. The relatively low gradient suggests the underlying mapping is not overly complex.

In order to assess the usefulness of the GA approach, the results for the generated sub-models will be compared with the results for the whole data-set. Table 6.2 through to Table 6.10 present the results for the sub-models.

Table 6.2 - Results for Residents Age Analysis

	MSE	R^2	SSq	COV	>15%
AGE1	14.0 [24.7]	0.90 [0.47]	65 [398]	753 [410]	45 [48]

The Genetic Algorithm was trained using Census variables describing the percentage of people in the surrounding neighbourhood falling into various age bands.

Table 6.3 - Results for House Type Analysis

	MSE	R^2	SSq	COV	>15%
HT1	16.6 [22.6]	0.85 [0.61]	64 [118]	433 [346]	38 [43]
HT2	14.8 [23.8]	0.90 [0.49]	68 [389]	703 [384]	36 [47]

The Genetic Algorithm was trained using Census variables describing the number of terraced, detached, flats, bedsits and semi-detached properties in the surrounding neighbourhood.

Table 6.4 - Results for Profession Analysis

	MSE	R^2	SSq	COV	>15%
PROF1	17.7 [19.9]	0.65 [0.47]	214 [286]	477 [376]	40 [47]
PROF2	18.3 [19.7]	0.69 [0.49]	215 [280]	490 [387]	38 [46]

The Genetic Algorithm was trained using Census variables describing the percentage of households where the head of the household is a manager, manual worker etc. in the surrounding neighbourhood.

Table 6.5 - Results for Car Availability Analysis

	MSE	R^2	SSq	COV	>15%
CAR1	15.5 [19.8]	0.78 [0.57]	64 [169]	311 [218]	32 [45]
CAR2	15.6 [19.4]	0.77 [0.57]	94 [161]	344 [266]	33 [44]
CAR3	12.2 [18.1]	0.84 [0.91]	323 [578]	169 [163]	30 [41]

The Genetic Algorithm was trained using Census variables describing the percentage of families without a car, owning 1 car, owning 2 cars etc. in the surrounding neighbourhood.

Table 6.6 - Results for Tenure Analysis

	MSE	R^2	SSq	COV	>15%
TEN1	14.7 [17.6]	0.80 [0.74]	675 [964]	231 [246]	29 [39]
TEN2	13.1 [19.8]	0.88 [0.66]	45 [147]	445 [339]	27 [42]

The Genetic Algorithm was trained using Census variables describing the percentage of homes that were owner outright, owner buying, rented from county councils etc. in the surrounding neighbourhood.

Table 6.7 - Results for Working Parents Analysis

	MSE	R^2	SSq	COV	>15%
WP1	21.0 [22.7]	0.55 [0.47]	234 [235]	349 [163]	42 [43]

The Genetic Algorithm was trained using Census variables describing the percentage of families that have both parents working, single parents working etc. in the surrounding neighbourhood.

Table 6.8 - Results for Ethnic Analysis

	MSE	R^2	SSq	COV	>15%
ETH1	19.7 [20.2]	0.61 [0.60]	452 [627]	425 [342]	40 [45]
ETH2	18.4 [19.4]	0.58 [0.59]	621 [349]	372 [401]	39 [44]

The Genetic Algorithm was trained using Census variables describing the percentage of families from various ethnic origins in the surrounding neighbourhood.

Table 6.9 - Results for Amenities Analysis

	MSE	R^2	SSq	COV	>15%
AMEN1	12.6 [19.2]	0.89 [0.38]	64 [385]	834 [442]	29 [45]
AMEN2	21.7 [21.9]	0.58 [0.51]	296 [308]	430 [398]	44 [45]
AMEN3	15.1 [17.7]	0.72 [0.66]	938 [119]	175 [256]	33 [44]

The Genetic Algorithm was trained using Census variables describing the percentage of homes without certain amenities such as an inside WC in the surrounding neighbourhood.

Table 6.10 - Results for Education Analysis

	MSE	R^2	SSq	COV	>15%
EDU1	12.5 [19.8]	0.90 [0.66]	41 [146]	447 [335]	29 [45]

The Genetic Algorithm was trained using Census variables describing the percentage of families with degrees and other qualifications in the surrounding neighbourhood.

Discussion of Single Category Results

To determine whether the sub-model approach has been successful for the single category analysis, it is useful to revisit the results. Table 6.11 presents the average sub-model error compared with the error observed for the single control model.

Table 6.11 - Summary of Results for Single Census Category Sub-Models

Census Category	Single Model	Sub Model	Improvement
Residents Age	22.2%	16.3%	5.9%
Economic Position	20.7%	20.5%	0.2%
Amenities	21.27%	19.9%	1.37%
Car Availability	19.9%	14.8%	5.1%
Tenure	23.2%	19.0%	4.2%
Working Parents	20.1%	15.6%	4.5%
Profession	22.9%	14.2%	8.7%
House Type	21.7%	15.1%	6.6%
Ethnic	21.5%	21.3%	0.2%
Migration	20.2%	20.6%	-0.4%
Travel to Work	23.04%	23.56%	-0.52%

Clearly, some of the sub-models outperformed the single model by a significant margin, whilst others showed comparable results. The results indicate an overall improvement in modelling accuracy for the sub-model approach compared to the single-model approach. The largest individual improvements are observed for: Profession; House Type; Car Availability; and Resident's Age models. Significant improvements were also made in Tenure and Working Parent sub-models.

Although this type of selective analysis is of some benefit in ascertaining underlying model parameters, it is more probable that a neighbourhood description encompasses more than just one geodemographic feature grouping. Table 6.12 presents results obtained for the GA analysis using a selection of those Census variables that proved to be beneficial in the single category analysis (See Appendix 1.2).

Table 6.12 - Results for a Selection of Mixed Census Variables

	MSE	R^2	SSq	COV	>15%
MIXED1	18.7 [19.5]	0.69 [0.51]	190 [218]	382 [318]	39 [43]
MIXED2	14.5 [28.4]	0.93 [0.74]	51 [205]	555 [568]	29 [54]
MIXED3	15.5 [18.1]	0.73 [0.69	82 [126]	201 [224]	32 [42]
MIXED4	16.0 [30.1]	0.92 [0.52]	57 [348]	801 [483]	32 [55]
MIXED5	15.9 [29.2]	0.92 [0.53]	55 [336]	437 [455]	29 [45]
MIXED6	18.9 [20.2]	0.60 [0.53]	209 [250]	319 [279]	38 [43]

The Genetic Algorithm was trained using a number of Census variables used in the single Census category analysis, including profession, tenure, residents age and car availability.

Discussion of Results for Multi-Category Census Data

Overall, the predictions made by the Census analysis sub-models outperformed those of single control model with an average increase in accuracy of 7.7%.

To appreciate the composition of the sub-models, it is useful to examine the individual chromosomes that define each sub-model and to decode them back into Census aggregates:

Sub-Model 1

Percentage of population unemployed is in the range 0 - 2%
Ratio of rooms per person is in the range 2 - 4 rooms per person
Percentage of mortgaged properties is in the range 0 - 6%
Percentage of local authority rented properties is in the range 0 - 9%
Percentage of semi detached properties is in the range 0 - 7%

Comparing the ranges contained in this sub-model with the whole data set, this sub-model describes an area with: low unemployment; high rooms per person ratio; low number of mortgaged properties; low number of local authority rented properties; and, low number of semi-detached properties.

Sub-Model 2

Ratio of rooms per person is in the range 2 - 4 rooms per person
Percentage of mortgaged properties is in the range 5 - 12%
Percentage of semi-detached properties is in the range 0 - 7%
Percentage of terraced properties is in the range 6 - 13%

This sub-model describes an area with: high rooms per person ratio; an average number of mortgaged properties; low number of semi-detached properties; and, an average number of terraced properties.

Sub-Model 3

Ratio of cars per person is in the range 0.3 - 0.6 cars per person
Ratio of rooms per person is in the range 2 - 4 rooms per person
Percentage of mortgaged properties is in the range 0 - 6%
Percentage of local authority rented properties is in the range 0 - 9%

This sub-model describes an area with: an average ratio of cars per person; a high ratio of rooms per person; a low number of mortgaged properties; and, a low number of properties rented from a local authority.

Sub-Model 4

Ratio of cars per person is in the range 0.3 - 0.6 cars per person
Percentage of mortgaged properties is in the range 0 - 6%
Percentage of local authority rented properties is in the range 0 - 9%

This sub-model describes an area with: an average ratio of cars per person; a low number of mortgaged properties; and, a low number of properties rented from a local authority.

Sub-Model 5

Ratio of cars per person is in the range 0.3 - 0.6%
Percentage of local authority rented properties is in the range 0 - 9%
Percentage of terraced properties is in the range 0 - 6%

This sub-model describes an area with: an average ratio of cars per person; a low number of properties rented from a local authority; and, a low number of terraced properties.

Sub-Model 6

Percentage population unemployed is in the range 0 - 2%
Ratio of rooms per person is in the range 2 - 4 rooms per person
Percentage of local authority rented properties is in the range 0 - 9%
Percentage of semi detached properties is in the range 6 - 13%

This sub-model describes an area with: a low number of unemployed people; a high ratio of rooms per person; a low number of properties rented from a local authority; and, an average number of semi-detached properties.

6.7 Analysis of Methodology

Before drawing any conclusions based on the described results, there are a number of issues relating to the method that require further discussion: initial population; Gamma Test; final population

Initial Population

This subject was broached earlier, where 3 approaches to generating the initial population were defined: random; dual mode; non-random. As previously anticipated, the most effective approach was the dual-mode with convergence significantly faster than using the random approach. Also the expected deterioration of the dual-mode approach with large numbers of Census aggregates was observed.

The third approach - the non-random approach - was used for data sets containing large numbers of Census aggregates (>15). However, this required a high mutation probability in order for the GA to diverge from the initial non-random generation. This in turn led to an unstable GA system with 'good' chromosomes being too heavily mutated. This may be improved by using a sliding scale of mutation which forces the first few generations to rapidly diverge from the initial population and then as the mutation probability decreases a stable population is reached.

Gamma Test

The Gamma Test is an important feature of this method as it provides the GA with its direction. On inspection of the results presented in Table 2 through to Table 14, it is evident that neither the Gamma intercept nor the Gamma gradient can be singly used as the fitness test. Rather, the estimation of trainability is best made using both. Unfortunately, the relationship between the Gamma metrics and the training accuracy is not direct as other issues effect the predictive accuracy such as: number of hidden layers; training time; sample size (overtraining) etc. The GA was directed using a weighted summation of a transformed Gamma intercept, Gamma gradient and sample size as defined in (3).

$$Fitness = \alpha * f(c) + \beta * g(m) + \chi * h(n) \tag{3}$$

where:

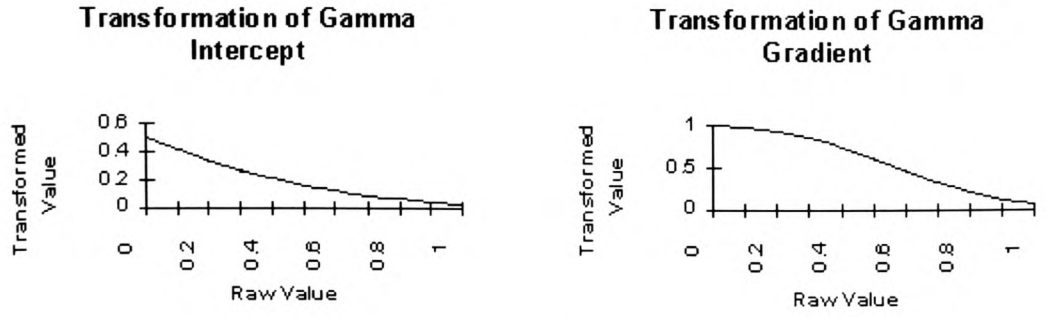
α: weight for intercept
β: weight for gradient
χ: weight for sample size

The transformations were those suggested by the authors of the Gamma test in the software documentation (Waggert, 1997).

$$f(c) = \frac{1}{(1 + \exp(\frac{c}{T}))} \quad \text{where T} = 0.3 \tag{4}$$

$$g(m) = \frac{2}{(1 + \exp(\frac{m^2}{T}))} \quad \text{where T} = 0.3 \tag{5}$$

Figure 6.7a and Figure 6.7b show the transformation of the raw Gamma metrics to their transformed values.



(a) *(b)*

Figure 6.7 - Transformation of the raw Gamma Metrics
(a) Transformation of Gamma Intercept.
(b) Transformation of Gamma Gradient

To direct the GA towards reasonable sample sizes, the sample size also forms part of the fitness function. This is defined as piecewise function as in (6).

if $n < md$

$$h(n) = \frac{n}{md} \tag{6}$$

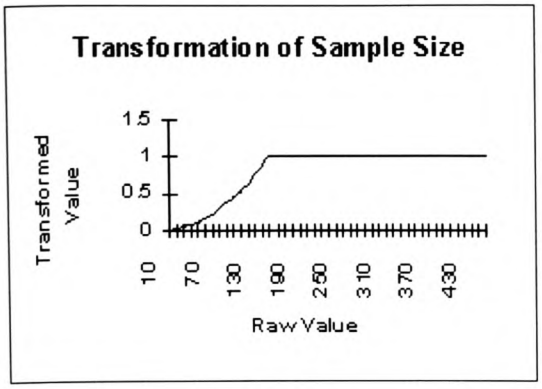
Otherwise

$$1$$

where

m : size of whole dataset
n : size of sample
d : percentage threshold

This piecewise approach encourages the GA to select chromosomes with larger sample sizes. The percentage threshold ensures that the sample size does not become overbearing in determining a chromosomes fitness. Figure 6.8 provides a graphical interpretation of the sample size part of the fitness function.



if n < md (7)
1
Otherwise

Figure 6.8 - Example Transformation of Sample Size.

Typical values for α, β, χ of 1, 0.5 and 0.1 were used for the described analysis.

Final Population

The population used to define the sub-models was composed of the best chromosomes encountered during the GA run. This proved to be more effective than just using the final population as on occasions, due to recent mutations, 'good' chromosomes were not present in the final population. During the GA run the top 10 chromosomes were recorded in an elite population.

6.8 Summary and Conclusions

To overcome the problem of post clustering fitness estimation, encountered using the clustering approach detailed in Chapter 5, the stratification problem was redefined using state space representation. This type of approach requires a method of navigation through the state space and a measure of fitness that can be applied to all states to determine direction:

Navigation: The chosen method for navigating the state space was a genetic algorithm. The GA approach is a non-exhaustive and non-optimal one, most suited to large state space problems. Binary coding was achieved by first discretising the raw Census data and then thresholding the discrete data over a number of soft partitions. Methods for generating initial populations were also investigated.

Fitness: The fitness function was a composite based mainly on the Gamma test. The Gamma intercept (noise) and Gamma gradient (complexity) were transformed to

fit a maximisation problem and combined with a thresholded sample size. Fitness was estimated for all chromosomes in a generation with the 'fittest' individuals propagating to successive generations.

MLP Sub-Models: An elite population was maintained containing the fittest individuals in a GA run. The elite population was used to stratify the whole data set into training sub-sets for independent MLP networks.

Results: The results show an increase in accuracy for the sub-models compared to a single control model. Moreover, the results are comparable with those obtained using the clustering approach with the added advantage of omitting the large redundancy found using the clustering method.

In conclusion, the search method is more appropriate when relationships between the stratification data and the underlying functionality of the training data are not sufficiently well understood. However, it must also be concluded that when these relationships are understood, it is easier to configure the clustering method as a user can simply make use of existing algorithms supplied with most neural network packages.

7. CONFIDENCE THROUGH COMPREHENSION

This chapter considers techniques that allow comprehensible rules to be extracted from neural network models. Firstly rules that describe the sub-markets proposed by the Kohenen Map are extracted. This is then repeated for the Genetic Algorithm Approach. These rules can be used to decide which MLP network to use when performing day-to-day valuations. Finally existing methods for extracting rules from MLP networks are considered and their use to valuers discussed.

7.1 Introduction

One of the major criticisms of neural networks that discourage acceptance by professionals, including residential valuers, is their inherent 'black box' nature. The internal workings of a neural network are, for the most part, hidden from the user. For a residential valuer, this leads to a situation where one problem is solved only for another to be created; a house may be valued accurately using neural network techniques but there is no explanation as to how the value was deduced, thus the valuation is not defensible in a court of law or valuation tribunal. One of the aims in building an intelligent system is to make neural network models more 'human-readable'.

The ability to furnish users with explanations of the reasoning process or underlying functionality is an important feature of any model (Clancy, 1983). Explanation facilities are required both for user acceptance and the validation of reasoning procedure (Davis, et al, 1977). This is a difficult task when analysing neural networks as they do not have explicit or *"declarative knowledge"* (Diederich, 1989). As a result of the numeric and distributed nature of neural networks, any rules extracted directly from their internal structure are often *"unintelligible to the human"* (Tay and Ho, 1992).

Acceptability of any system requires that there be a degree of understanding of the system by its operators. The level of understanding is more closely analogous - in

relation to a car - to the driver than the mechanic, though some users will become 'mechanics'. If an intelligent system is to be used by appraisal professionals, they must have a reasonable understanding of its processes, be confident in explaining them and recognise when there is something wrong.

Expert systems are sufficient in this respect. A true expert system is always able to explain its decision process. Neural networks, however, are as much of a 'black box' as DCC. To be acceptable, the neural network component of any intelligent system needs to be transparent. This is a challenge that cannot be avoided.

"The conventional approach to building an expert system requires a human expert to formulate the rules by which the data can be analysed" (Zurada, 1992).

In contrast, a connectionist or induction expert system formulates its knowledge base by modelling implicit functions or relationships within a data-set. A connectionist expert system is in essence a straightforward neural network. However, the environment software examines the weights recorded for a trained network and attempts to give a tentative explanation of the model based on the magnitude of the weights. Unfortunately, the explanations generated by a connectionist expert system are far removed from the symbolic heuristics taken by a human expert and more in tune with the synoptic activity of the brain. It has been suggested that the *"one thing that connectionist networks have in common with brains is that if you open them up and peer inside, all you can see is a big pile of goo"* Mozer and Smolensky (1989). However, despite this complexity, many researchers are investigating methods of extracting rules from trained networks (see Andrews, et al, 1995 for an introduction to rule extraction techniques).

Understanding breeds confidence. Those who use the systems will need to be confident in their use. They need to know the limits to which the system may be safely driven. They will also need reassurance about system suitability when the terrain changes. For this reason, the intelligible system will also need to provide the user with information about the degree of confidence that is appropriate to any current transaction. Intelligibility requires that even if the calculation of confidence is complex, the communication of confidence to the user is simple to understand.

7.2 Rule Extraction

One approach to understanding a neural network model is to translate the hypothesis represented by a trained neural network into a more comprehensible language, namely an inference-rule language. This task is defined as:

"Given a trained neural network and the data on which it was trained, produce a description of the network's hypothesis that is comprehensible yet closely approximates the network's behaviour" (Craven, 1996). Such systems learn rules from *"raw domain data"* (Quinlan, 1986; Michalski and Chilansky, 1980).

Rule extraction should address the following four criteria (Craven, 1996):

Comprehensibility: Symbolic representations of neural network hypothesis should be comprehensible to experts in the domain.
Fidelity: Symbolic representations of neural network hypothesis should accurately model and embody the underlying functionality of the networks they were extracted from.
Scalability: Symbolic representations should be scalable for all possible network sizes.
Generality: Rule extraction should not impose special training regimes or restrictions to network architecture.

Rule Extraction Via Sub-Market Definition

Given the problems associated with direct rule extraction from neural network structure, investigation focused on identifying rules that describe the decisions made by the Kohonen network as opposed to the predictions made by the back propagation network. By examining the way the modular system works, it is evident that each sub-model (which in essence represent a homogeneous data set) is constructed using a subset of the features present in the parent data set. Moreover, this subset of features differs from one sub-model to the next. Therefore, rules extracted by inspection of the feature subset describe which sub-model can best predict the value of a previously unseen property.

Property Type: The general mix of housing stock in a region may have an impact on property prices. Statistics are available on detached; semi-detached; terraced; flats; and bed-sits.

The following rules were extracted from the Kohonen SOM trained on property type data.

If Detached-Properties in Neighbourhood is between 3 and 20%
And Semi-Detached-Properties in Neighbourhood is between 67 and 85%
And Terraced-Properties in Neighbourhood is between 1 and 15%
And Purpose-Built-Flats in Neighbourhood is between 0 and 5 %
And Converted-Flats in Neighbourhood is 0%
And Bedsits in Neighbourhood is 0%
Then Network-Selection is 'HTYPE2'

Figure 7.1 provides a graphical representation of the profile of the records used to train network HTYPE2.

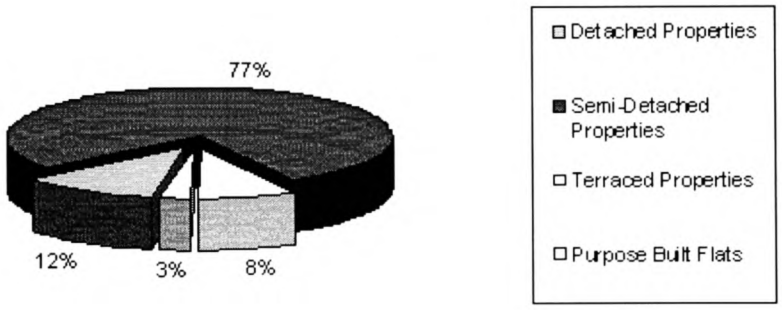


Figure 7.1 - Profile of Records used to Train Network HTYPE2

The neighbourhood from which the training records were selected for network HTYPE2 has a high proportion of semi-detached properties. Interestingly, the next largest housing stock category is detached houses - suggesting this is a probably a suburban area.

If Detached-Properties in Neighbourhood is between 0 and 4%
And Semi-Detached-Properties in Neighbourhood is between 2 and 12%
And Terraced-Properties in Neighbourhood is between 74 and 94%
And Purpose-Built-Flats in Neighbourhood is between 0 and 13 %
And Converted-Flats in Neighbourhood is 0 and 2%
And Bedsits in Neighbourhood is 0%
Then Network-Selection is 'HTYPE1'

Figure 7.2 provides a graphical representation of the profile of the records used to train network HTYPE1.

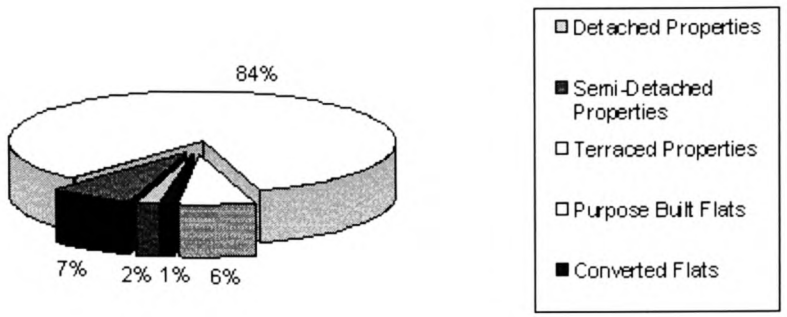


Figure 7.2 - Profile of Records used to Train Network HTYPE1

The neighbourhood from which the training records were selected for network HTYPE1 has a high proportion of terraced properties. The remaining housing stock is mainly made up from semi-detached properties and purpose built flats.

If Detached-Properties in Neighbourhood is between 3 and 35%
And Semi-Detached-Properties in Neighbourhood is between 12 and 29%
And Terraced-Properties in Neighbourhood is between 22 and 49%
And Purpose-Built-Flats in Neighbourhood is between 4 and 40%
And Converted-Flats in Neighbourhood is between 0 and 1%
And Bedsits in Neighbourhood is 0%
Then Network-Selection is 'HTYPE4'

Figure 7.3 provides a graphical representation of the profile of the records used to train network HTYPE4.

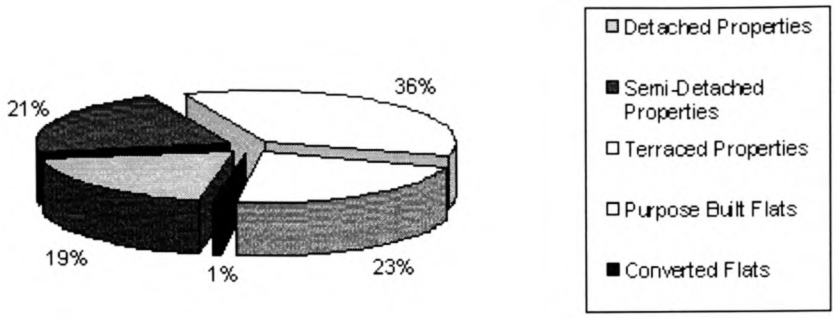


Figure 7.3 - Profile of Records used to Train Network HTYPE4

The neighbourhood from which the training records were selected for network HTYPE4 is heterogeneous with respect to housing stock.

Ethnicity: The ethnic make up of an area may have an influence on property value. However, Openshaw and Wymer (1994) warn of the danger of associating multi-ethnic areas with poorer areas and suggests that tenure is used to distinguish between *"financial stable and financially stressed multi-ethnic regions"* (Openshaw and Wymer, 1994).

The Kohonen map was unable to discern any discrete clusters using ethnicity data, this was because of the monolithic nature of the data with 96% of the selected population classed as White-European.

Tenure: Statistics describing nine tenures are available: Owner Occupied Outright; Mortgaged; Privately Rented (Furnished); Privately rented (Unfurnished); Rented with Business; Local Authority Rented; Housing Association Rented; Armed Forces Rented. Openshaw and Wymer (1994) state *"Generally, the rented category and especially accommodation rented from Local Authorities etc. has been used by researchers as a measure of lack of resources and residential insecurity. In contrast, because of the financial commitment required to purchase a house, house ownership is seen as a surrogate for long term financial stability."*

The following rules were extracted from a Kohonen SOM trained using statistics on residents' tenure.

If Percentage-Owner-Occupied (Outright) is between 43 and 45%
And Percentage-Owner-Occupied (Buying) is between 32 and 40%
And Percentage-Privately-Rented (Furnished) is between 1 and 3%
And Percentage-Privately-Rented (Unfurnished) is between 2 and 6%
And Percentage-Rented-From-Housing-Association is between 0 and 2%
And Percentage-Rented-From-Local-Authority is between 0 and 6%
Then Network-Selection is 'Tenure3'

Figure 7.4 provides a graphical representation of the profile of the records used to train network Tenure3.

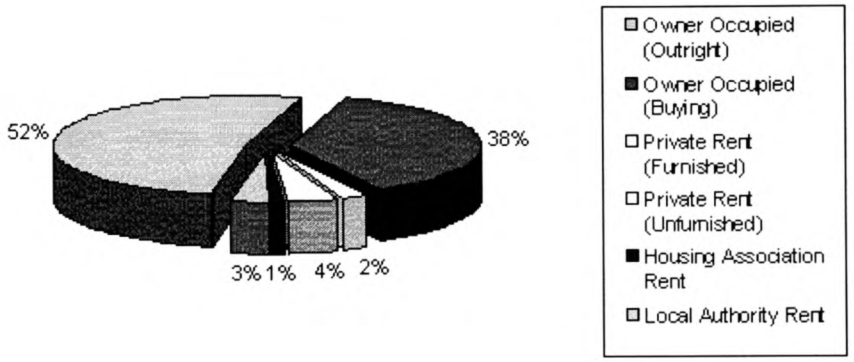


Figure 7.4 - Profile of Records used to Train Network Tenure3

The neighbourhood from which the training records were selected for network TENURE3 has a high proportion (90%) of owner occupied properties - with an average of 38% mortgaged.

Car Availability: Openshaw and Wymer (1994) used the percentage of households lacking a car as a measure of short-term financial deprivation. Statistics are available on those households with: no car; 1 car; 2 cars; 3+ cars.

The following rules were extracted from a Kohonen SOM trained using statistics on availability of cars.

IF Percentage-Households-With-No-Car is between 34 and 42%
And Percentage-Households-with-1Car is between 40 and 44%
And Percentage-Households with 2Cars is between 14 and 19%
And Percentage-Households with 3+Cars is between 2 and 5%
Then Network-Selection is 'Cars2'

Figure 7.5 provides a graphical representation of the profile of the records used to train network Cars2.

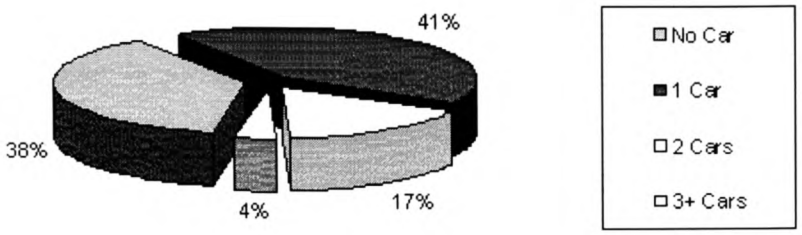


Figure 7.5 - Profile of Records used to Train Network CARS2

The neighbourhood from which the training records were selected for network CARS2 has a high proportion of households with 2 cars.

IF Percentage-Households-With-No-Car is between 26 and 35%
And Percentage-Households-with-1Car is between 49 and 55%
And Percentage-Households with 2Cars is between 12 and 17%
And Percentage-Households with 3+Cars is between 2 and 4%
Then Network-Selection is 'Cars4'

Figure &.6 provides a graphical representation of the profile of the records used to train network Cars4.

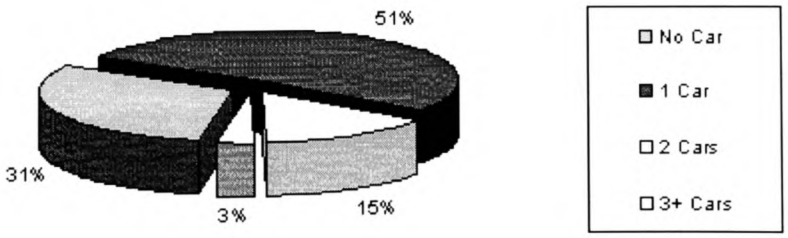


Figure 7.6 - Profile of Records used to Train Network CARS4

The neighbourhood from which the training records were selected for network CARS4 has mainly between 0 and 1 cars per household.

Socio-Economic: Census variables describing the proportion of people in an area who are economically active (employers, self-employed, employees etc.), their skills in an industry (managerial, professional, semi-skilled etc.) and also the type of industry (agricultural, manufacturing etc.) are available. Statistics describing workers qualifications (higher degrees, diplomas etc.) are also available. Finally, the proportion of unemployed people in a community can be estimated using Census statistics.

The following rules were extracted from a Kohonen SOM trained using statistics on residents' qualifications.

If Qualified-Persons is between 25 and 34%
And Higher-Degree-Qualifications is between 0 and 3%
And Degree-Qualifications is between 8 and 18%
And Diploma-Qualifications is between 13 and 19%
Then Network-Selection is 'EDU2'

If Qualified-Persons is between 12 and 16%
And Degree-Qualifications is between 2 and 7%
And Diploma-Qualifications is between 7 and 11%
Then Network-Selection is 'EDU4A'

The following rules were extracted from a Kohonen SOM trained using statistics on residents' socio-economic standing.

If Percentage-Full-Time-Employment is between 30 and 37%
And Percentage-Part-Time-Employment is between 8 and 10%
And Percentage-Self-Employed (with Employees) is between 1 and 6%
And Percentage-On-Government-Schemes is between 0 and 1 %
And Percentage-Unemployed is between 3 and 8%
Then Network-Selection is 'Employ1'

Figure 7.7 provides a graphical representation of the profile of the records used to train network Employ1.

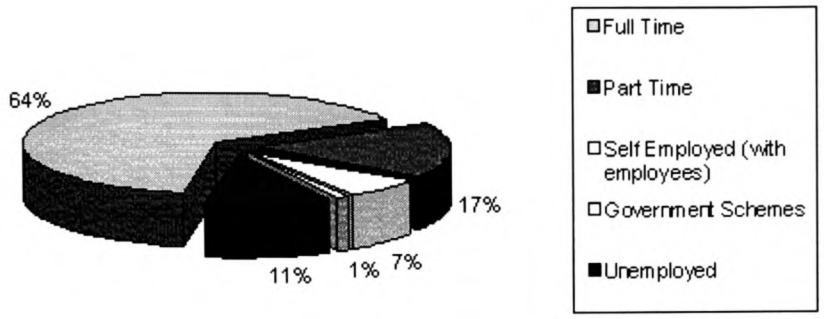


Figure 7.7 - Profile of Records used to Train Network Employ1

The neighbourhood from which the training records were selected for network Employ1 has a high proportion of residents who are full-time employed (30-37% of all residents or an average of 64% of those aged 16-65).

7.3 Rules Extraction from the Search Stratification Algorithm

Rule extraction for the search algorithm approach simply involves decoding the individual chromosomes that define each sub-model into Census aggregates:

Sub-Model 1

If Percentage of population unemployed is in the range 0 - 2%
And Ratio of rooms per person is in the range 2 - 4 rooms per person
And Percentage of mortgaged properties is in the range 0 - 6%
And Percentage of local authority rented properties is in the range 0 - 9%
And Percentage of semi detached properties is in the range 0 - 7%
Then Network-Selection is 'GA1'

Comparing the ranges contained in this sub-model with the whole data set, this sub-model describes an area with: low unemployment; high rooms per person ratio; low number of mortgaged properties; low number of local authority rented properties; and, low number of semi-detached properties.

Sub-Model 2

If Ratio of rooms per person is in the range 2 - 4 rooms per person
And Percentage of mortgaged properties is in the range 5 - 12%
And Percentage of semi-detached properties is in the range 0 - 7%
And Percentage of terraced properties is in the range 6 - 13%
Then Network-Selection is 'GA2'

This sub-model describes an area with: high rooms per person ratio; an average number of mortgaged properties; low number of semi-detached properties; and, an average number of terraced properties.

Sub-Model 3

If Ratio of cars per person is in the range 0.3 - 0.6 cars per person
And Ratio of rooms per person is in the range 2 - 4 rooms per person
And Percentage of mortgaged properties is in the range 0 - 6%
And Percentage of local authority rented properties is in the range 0 - 9%
Then Network-Selection is 'GA3'

This sub-model describes an area with: an average ratio of cars per person; a high ratio of rooms per person; a low number of mortgaged properties; and, a low number of properties rented from a local authority.

Sub-Model 4

If Ratio of cars per person is in the range 0.3 - 0.6 cars per person
And Percentage of mortgaged properties is in the range 0 - 6%
And Percentage of local authority rented properties is in the range 0 - 9%
Then Network-Selection is 'GA4'

This sub-model describes an area with: an average ratio of cars per person; a low number of mortgaged properties; and, a low number of properties rented from a local authority.

Sub-Model 5

If Ratio of cars per person is in the range 0.3 - 0.6%
And Percentage of local authority rented properties is in the range 0 - 9%
And Percentage of terraced properties is in the range 0 - 6%
Then Network-Selection is 'GA5'

This sub-model describes an area with: an average ratio of cars per person; a low number of properties rented from a local authority; and, a low number of terraced properties.

Sub-Model 6

If Percentage population unemployed is in the range 0 - 2%
And Ratio of rooms per person is in the range 2 - 4 rooms per person
And Percentage of local authority rented properties is in the range 0 - 9%
And Percentage of semi detached properties is in the range 6 - 13%
Then Network-Selection ig 'GA6'

This sub-model describes an area with: a low number of unemployed people; a high ratio of rooms per person; a low number of properties rented from a local authority; and, an average number of semi-detached properties.

7.4 Conclusions

Acceptability of any system requires some understanding of the system by its operators. If an intelligent system is to be used by appraisal professionals, they must have a reasonable understanding of its processes, be confident in explaining them and recognise when there is something wrong.

The rules described in this chapter are of a form that are easily understandable and relate directly to the type of market the training records are selected from. This type of approach will allow the user to make judgement as to the suitability of the training set as a base for which to estimate the value of a subject property and, to use his experience in assessing whether the estimated value is within a reasonable bounding.

The rules extracted from these complex distributed models are able to furnish the user with a reasonable level of understanding, sufficient to allow him to make valued judgement and hence it can be concluded that this provides the transparent 'readable' property required of intelligent appraisal models.

8. CONCLUSIONS AND FUTURE WORK

This chapter summarises the research documented in this thesis and brings together the conclusions reached during this project. Overall conclusions and limitations are discussed which lead to suggestions for future work.

8.1 Introduction

At the start of this thesis, the aims of the research were detailed, with the general aim of the research being to assess the potential of AI techniques as tools to assist in the appraisal of residential properties. In particular the research sought to:

- Assess empirically the suitability of Artificial Neural Network models to assist a residential property valuer in day-to-day valuations;
- Use existing or develop new techniques that facilitate cognitive understanding of the underlying reasoning processes of ANN models;
- Specify a prototype system that could be 'bolted' onto a comparables database to provide ANN estimates of value.

The empirical research presented in the preceding chapters provide a means of addressing these research objectives, with this Chapter bringing together the fore-mentioned techniques and suggesting approaches for integrating these novel techniques into an appraisal system.

8.2 Previous Related Work

Interest in residential property research comes, firstly, from the consequences of the property crash in the late 1980s and, secondly, from the desire of the valuation profession to utilise new technologies and emerging IT techniques. Literature refers

to a number of alternative techniques for building appraisal models from transaction data, of which the following were the most frequently mentioned:

- Multiple Regression Analysis (MRA)
- Linear Programming (LP)
- Artificial Neural Networks (ANNs)
- Expert Systems and Expert Database (ES and ED)

MRA is the most commonly used technique. However, the drawbacks of this technique are its restriction to modelling only linear relationships and the danger of reducing modelling ability from unresolved skewness and multi-colinearity.

Linear programming is a technique proposed only by Wiltshaw (1991a,1991b,1993) as a mathematical approach to modelling residential property values. This technique is based on firm mathematical foundations. However, this same mathematical structure poses a restriction on the application of this technique due to its algebraic requirements.

Next to MRA, artificial neural networks are the most commonly used technique within published literature. Their ability to easily model non-linear relationships - that are instantly recognisable within residential property data - puts this technique ahead of MRA. However, the drawback of this technique is the inability of neural network models to express their underlying functional form.

In addition to these regression techniques, another popular line of research is the development of an expert system that replicates the actions of the professional valuer. This is perhaps the easiest of techniques to comprehend but also the most difficult to build. The 'experience factor' used by professional valuers is difficult to replicate in explicit rule based structures. However, it is proposed by some artificial intelligence researchers, that these techniques can be combined with regression techniques in a hybrid manner that compensate for the weaknesses in the individual techniques.

From the literature, it can be concluded that there is an abundance of research potential in the field of residential property appraisal, with many unanswered questions relating to MRA and neural networks.

Future research should concentrate on complementing the weakness of one method with the strength of another (Goonatilake and Khebbal, 1995) to provide the framework for a holistic model (Gronow, et al, 1996). However, it is noteworthy that the systems being developed are to act as instruments rather than appraisers. Researchers, in the most part, agree that *"Like an automatic pilot, the real decisions have to be taken by the professional who acts after the instruments have produced the basic information on current conditions"*. (James, 1994)

8.3 Assess empirically the suitability of ANN models to assist a residential property valuer in day-to-day valuations

To assess the potential ANNs have as appraisal models, empirical research began by building a simple ANN appraisal model as proposed in recent research publications (Evans, et al, 1992; Borst, 1991).

Conventional ANN Model

The conventional approach to building a neural network appraisal model was demonstrated using training data selected from a homogeneous region. The aim of the exercise was to assess the validity of the data to act as training set prior to pursuing the research objectives. Validity was measured by comparing models trained using the data against published benchmarks. The data was shown to be useful as the training set for a neural network model, given the selected data came from a relatively homogeneous area.

An application for this simple approach is suggested as being the estimation of property value for taxation purposes. As in the previous example, a neural network model was trained with a single output node representing value, this was compared with a further neural network model representing value as a banded classification. An additional model based on Quinlan's tree induction techniques was also investigated. The results show that Quinlan's model performed significantly better than the neural network models when predicting examples from the training set,

however, it was the neural network models that came to the forefront when predicting previously unseen examples due to their generalisation qualities.

However, when this approach was used to model data extracted from a more heterogeneous area with respect to value, it soon became clear that predictive accuracy decreased. The research confirms the hypothesis that location has a major influence on the value of a residential property.

Modelling Data Selected from an Heterogeneous Area

Literature suggests that location is the primary influence on residential property value (Mackmin, 1994). However, location in itself is not something that can be strictly defined as can the number of bedrooms or the existence of a garage.

Location is subjective, it changes over time and is often unique. Professional appraisers must spend time getting accustomed to an unfamiliar location in order to appreciate and correctly interpret all aspects of a society (Mackmin, 1994). It is without doubt, therefore, that qualitative measure of location, neighbourhood and local economies form an integral part of any computer assisted residential appraisal model.

The average values of similar properties in a neighbourhood would seem to be a good place to start. Indeed, the empirical evidence presented in Chapter 4 supports this approach. However, to gain a good measure of value ranges in an area requires a good cross-section of representative data. Furthermore, the average values only really make sense if they are constructed for similar types of properties. This method then becomes similar to the DCC method that it is trying to support and suffers from the same problems of data scarcity etc.

A description of a location's wealth, amenities, housing stock etc. could be a useful addition to an appraisal model. This level of information is available in the form of raw Census data and summarised Census data (geodemographic systems). The inclusion of Census data into a computer assisted appraisal model improved the accuracy of the model by an average of 2% using data at the District level and 7% using data at Enumeration District level.

By accessing more dynamic sources of data and including subjectively assigned variables from professional valuers, it was believed that appraisal models could be significantly enhanced.

The research presented in Chapter 4 shows that the addition of Census data at the ED level into an ANN appraisal model significantly increased its accuracy. However, with consideration to a priori knowledge relating to the varied interplay of demand and supply side variables across different geographical regions, it was concluded that the Census data could be more effectively employed as a method of segregating the heterogeneous property market into homogeneous sub-markets.

8.4 Stratification using Clustering Techniques

The stratification approach was first investigated using just property data. This research led to the development of a novel technique, which takes heterogeneous training data and clusters it into homogeneous sub-sets using a Kohonen SOM. The results obtained for clusters found in property types, show that the methodology compares very favourably with the more conventional neural network approach. An average increase in prediction accuracy of 10% was achieved using the new method over the conventional approach. This implies that the original data set either contained more than one underlying function (pattern) (James, 1994) or the function was too elaborate to be modelled using a single back propagation network.

The technique employed to estimate the susceptibility of the data to be modelled - i.e. does a particular cluster contain useful training records? - was the Gamma test. Based on a nearest-neighbour approach, this method gives a measure of both noise (intercept) and complexity (gradient), assuming a single smooth continuous function underpins the data set. Two assumptions can therefore be made about a set of data given its Gamma results. Firstly, a data set with a high noise value may have insufficient examples, descriptive features or contain data mapped by multiple underlying functions. Secondly, a Gamma result showing a very complex underlying function may in fact be an aggregate of multiple functions that may be too complex for MLP or MRA to model.

Extending the Clustering Technique to Include Demand-Side Data

Given the encouraging results obtained using property data, the method was then used to cluster Census data in the hope that areas having similar Census characteristics would share similar valuation functions.

Census aggregates were passed as input vectors to a Kohonen SOM that clusters data according to their cross-characteristic similarities. After training, each cluster represented a set of residential properties linked via an enumeration to postcode cross-reference look up table. The Gamma test was used to estimate the trainability of each subset, with those with low noise and low complexity forming training sets for individual MLP networks. The results show that the accuracy of the sub-models outperformed a single MLP control model within the range of 1 to 14%, with an average increase in modelling accuracy of 5%.

Census clusters were used to describe the content of a collection of training sets that were each modelled independently using an MLP network. The outcome of this suggests that:

- A set of models, each dedicated to a certain narrow domain, can significantly outperform predictions made by a single more general model trained on all of the available training data.
- Models created from the stratification technique can be used to predict property values in other areas that have similar Census characteristics.

Although these results are promising, the generation of suitable training sets relies heavily upon selection of useful neighbourhood characteristics. Poor selection leads to clusters forming which perform badly when analysed by the Gamma test. This method is therefore of most use when a priori knowledge is available to determine the neighbourhood descriptors to select.

8.5 Stratification using Search Techniques

To overcome the problem of post clustering fitness estimation encountered using the clustering approach, the stratification problem was redefined using state space representation. This type of approach requires a method of navigation through the

state space and a measure of fitness that can be applied to all states to determine direction:

Navigation: The chosen method for navigating the state space was a genetic algorithm. The GA approach is a non-exhaustive and non-optimal one, most suited to large state space problems. Binary coding was achieved by first discretizing the raw Census data and then thresholding the discrete data over a number of soft partitions. Methods for generating initial populations were also investigated.

Fitness: The fitness function was a composite based mainly on the Gamma test. The Gamma intercept (noise) and Gamma gradient (complexity) were transformed to fit a maximisation problem and combined with a thresholded sample size. Fitness was estimated for all chromosomes in a generation with the 'fittest' individuals propagating to successive generations.

MLP Sub-Models: An elite population was maintained containing the fittest individuals in a GA run. The elite population was used to stratify the whole data set into training sub-sets for independent MLP networks.

The results show an increase in accuracy for the sub-models compared to a single control model. Moreover, the results are comparable with those obtained using the clustering approach with the added advantage of omitting the large redundancy found using the clustering method.

Here, the results are more consistent than the Kohonen approach as fitness is tested on-route as opposed to post-clustering. An average increase in accuracy of 7.5% was observed, with some models improving by up to 16%. Some models made no aggregate improvement over the single control model with a few models fairing marginally worse.

In conclusion, the search method is more appropriate when relationships between the stratification data and the underlying functionality of the training data are not sufficiently well understood. However, it must also be concluded that when these relationships are understood, it is easier to configure the clustering method as a user

can simply make use of existing algorithms supplied with most neural network packages.

8.6 Use existing or develop new techniques that facilitate cognitive understanding of the underlying reasoning processes of ANN models

Acceptability of any system requires some understanding of the system by its operators. If an intelligent system is to be used by appraisal professionals, they must have a reasonable understanding of its processes, be confident in explaining them and recognise when there is something wrong.

Expert systems are sufficient in this respect. A true expert system is always able to explain its decision process. Clearly, the ability to produce rules that provide a level of understanding of the functionality of the intelligent models will enhance its potential of being accepted in the valuation community.

The rules elicited from the Kohonen and Genetic Algorithm models are of a form that are easily understandable and allow the user to relate each model to the type of market from which the training records were selected. This ability of the model to expose its underlying functionality as a set of 'human readable' rules will increase the potential of such a model to be accepted within the professional valuation community.

The rules extracted from these complex distributed models are able to furnish the user with a reasonable level of understanding, sufficient to allow him to make valued judgement and hence it can be concluded that this provides the transparent 'readable' property required of intelligent appraisal models.

8.7 Specify a prototype system that could be 'bolted' onto a comparables database to provide ANN estimates of value

In order to appreciate how these research techniques could be integrated into a useable residential property appraisal system, previous work in the field of residential property appraisal undertaken by the University of Glamorgan needs to be introduced. Specifically, the development of an Expert Database that automates the comparable process making use of handheld personal computer technology to act as an electronic notebook linked to a database server that estimates value and

generates valuation reports. The system known as NIMROPA was developed as a research tool (a variation of the system was field tested by a leading UK lending institute). Figure 8.1 provides a schema of the hypothesised appraisal system and at a high level shows how these data techniques (e.g. neural networks) could be used to elicit functions and make estimations of value that could be used in addition to traditional (automated) comparable approach as embedded systems.

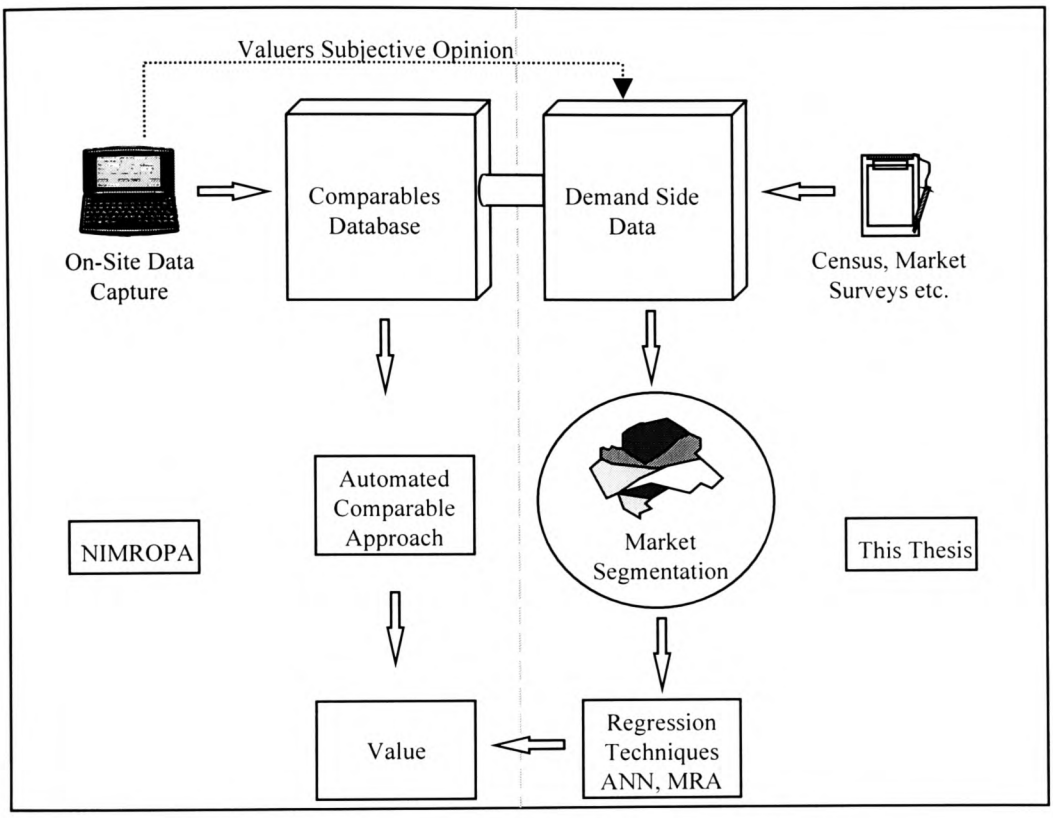


Figure 8.1 - Integration of Intelligent Model into a Hybrid Appraisal Model

Figure 8.1 presents a model based around a central resource of transaction data. This can be used to automate the comparable approach to residential property valuation (Gronow, et al, 1996). The hypothesis being that coupling the traditional approach with techniques investigated in the previously described research can enhance information available to a valuer. For instance, it is quite feasible that the genetic algorithm presented in Chapter 6 could be used to develop neural network models based on transaction and neighbourhood data and thus provide the professional valuer with an additional estimation of value for the subject property. This remains as an hypothesis within the context of this thesis and is drawn upon in the future work section.

Assessing Potential

The task of the valuer in the residential property market is a non-trivial one; many factors - some tangible and others intangible - come into play during an appraisal. This thesis has presented an analysis of how several different techniques can be used to assist in this task. What needs to be made clear is that none of the proposed methods individually offer an holistic solution. Rather, they provide tools to help in the task and should enable the valuer to take a more informed decision and thus provide consistent and accurate valuations.

To summarise the performance and merits of each of the techniques investigated in this research, a number of metrics commonly used in published literature have been grouped to form the following discussion categories:

Accuracy: The accuracy of a technique encapsulates both the modelling of homogeneous data and heterogeneous data. The following methods were used to measure the accuracy of the generated models in predicting the value of residential properties: R^2 (Adair, et al, 1996; Lam, 1996); MSE (Evans, et al, 1992; Worzala, et al, 1995; Lu and Lu, 1992); SSq (Lam, 1996); Gamma Test; and, Covariance (Steffanson, et al, 1997; Lam, 1996).

Automated Modelling: This measure relates to the ease at which a data-set containing mortgage transaction data selected from a heterogeneous region could be automatically used as a training set for a regression based modelling technique whilst still conserving accuracy levels achieved for homogeneous training data. The process of stratification or segmentation is often done manually (see Adair, and McGreal, 1995 stratification of Belfast). However, a number of researchers have proposed that this step should be automated prior to engaging in regression based modelling (Almy, et al, 1998, James, 1994)

Transparency: Here, quality is a measure of 'human readability' of the generated models. Good models are able to express their underlying functionality in a manner easily understood by their users. Many researchers cite neural networks as being deficient due their inability to express their underlying functional form a symptom often described as 'black-box' (Worzala, et al, 1995; McClusky and Adair, 1998)

Practicality: This measure makes a judgement of practicality for each of the investigated techniques based on properties highlighted within published literature and other supportive evidence. For each investigated technique, the following is discussed: ease of use; exposure to academic and valuation communities (based on novelty of each technique); support in commercially available applications; and, potential for inclusion in a valuation tool as an embedded modelling technique.

Conventional ANN Approach

The first technique that was investigated was the conventional ANN approach that replicated the work found in previous literature (see Evans, et al, 1992; Tay and Ho, 1992; Do and Grudnitski, 1992). As in the previous studies it was found that a good degree of accuracy was obtained for analysis based on data selected from homogeneous regions. However, attempting to generate accurate models based on data selected from heterogeneous regions proved to be non-productive. This method, therefore, scores low on the automated front as accuracy levels are only acceptable for homogeneous data.

The conventional ANN approach - based on the multi-layer perceptron trained using the back-propagation algorithm - is well documented both within academic and professional press with articles regularly appearing in the RICS conferences and major journals such as the Journal of Property Research.

The major deficiency cited for this technique being its inability to express its underlying functional form (Worzala, et al, 1995; McClusky and Adair, 1998). Although there are methods to extract production rules from trained MLP networks (see Andrews, et al, 1995 and Craven, 1996) these are often unintelligible to the average user (Tay and Ho, 1992) and therefore this criticism remains valid for the conventional ANN approach.

Quinlan's Tree Induction Approach

The second method investigated, albeit for the specific purpose of council tax valuations, was Quinlan's Tree Induction Method C5.0. The method, in its simplest form, branches for all classifications within the supplied training attributes. For example, a particular branch may represent all 3-bedroom homes, with gas central

heating, built prior to 1950. Because of the high correlation between the training set and the generated tree structure it was found that this technique predicted to a high level of accuracy those homes included in the training set but to a much lower level of accuracy for the holdout sample.

The models generated using Quinlan's C5.0 technique are highly transparent with rules similar to those used in expert systems produced as the model output. For each branch, a ratio is given that relates to the number of examples where their attributes matched those of the specified branch and their target class matched that of the branch leaf node class.

Based on the results gained, this method warrants further investigation for inclusion in an 'intelligent' appraisal model. Emphasis for further research should be placed on the creation of highly representative training sets that will facilitate a higher degree of accuracy for holdout samples.

There are a number of implementations of this technique available as research resources, however the method requires specific data transformations and code tweaks to suit the proposed problem. There appear to be no other attempts at using Quinlan's C5.0 for residential property valuation research in academic publications let alone professional press.

Cluster Stratification Approach using a Kohonen Self-Organising Map (SOM)

To overcome the failings of the previously investigated methods with respect to model accuracy for heterogenous data, a Kohonen SOM was used to find homogeneous clusters (subsets of the heterogeneous parent dataset) suitable as training sets for regression based models. In addition to using the Kohonen SOM, a recent innovative technique called the Gamma Test was used to estimate the suitability of each cluster as a training set and to allow 'useless' clusters to be rejected.

This hybrid method shows significant potential, with automated stratification of the heterogeneous data-set into homogeneous sub-sets using the Gamma test as a state evaluation function making the approach highly scaleable and increasing accuracy over the whole sample data. Overall accuracy levels - using the previously described

metrics - were shown to be higher for the most sub-models compared with the single conventional ANN approach. The method suffers however from a high generation of redundant clusters through stratification using non-influential variables and post-cluster fitness evaluation.

This approach could be replicated using most commercially available neural network software packages. However, the automation required for this research was achieved by the creation of additional software components to extract clusters and compute their Gamma metrics. Publications on the use of the Kohonen SOM are widely available as academic papers and books (see Bishop, 1994). However, the use of the Kohonen map in combination with the Gamma test for residential property appraisal analysis is novel and hence not impacted on professional press.

An additional benefit accrued from this method is the ability of 'human-readable' rules to be extracted describing the composition of the homogeneous subsets used as training sets for the regression based models. This therefore increases the potential of such a technique to be accepted within the valuation community.

The stratification method clearly showed significant potential to warrant further research into techniques that would alleviate the redundancy problem and requirement for well informed selection of variables to be used as the training set for the Kohonen model. This requirement led to the development of a genetic algorithm approach to the stratification stage of the appraisal model.

Search Stratification Approach using a Genetic Algorithm

The final method investigated was stratification using a genetic algorithm that allowed fitness of training sets to be estimated iteratively. The genetic algorithm - using Gamma metrics as its fitness function - progressed the training sets that showed most benefit and randomly generated new ones at each generation until an 'elite' population of training sets was achieved.

The accuracy of the models created using this method were at least as good as those produced by the Kohonen stratification process. Moreover, this method was significantly more automated and therefore having greater potential to form an embedded process within an intelligent appraisal system.

Some commercial neural network packages support genetic algorithms and therefore this method could be replicated using commercially available packages. However, to achieve suitable automation and 'ease of use' this technique would require some software development.

Again, this technique expresses the underlying composition of the homogeneous sub-sets by decoding the generated chromosomes and representing them in a familiar expert system rule base format.

8.8 Potential Impact of an Intelligent Residential Appraisal System

The development of an intelligent residential property appraisal system would have impact in the areas of mortgage valuation and mass appraisal for taxation purposes.

Clearly, for taxation purposes the development of an intelligent appraisal model is attractive due to the mammoth task involved in mass appraisal. McCluskey and Adair (1998) suggest that for mass appraisal systems, estimation of value should be based on homogeneous groups of properties.

The stratification methodologies presented in this thesis - particularly the genetic algorithm technique - lend themselves well to mass appraisal given their ability to segment heterogeneous data-sets containing multiple valuation functions into homogeneous sub-sets suitable as training sets for an ANN or MRA based regression model.

With regard to valuation for mortgage purposes, an intelligent appraisal model could complement the values gained from the conventional DCC approach. As well as providing valuers with additional evidence, an intelligent appraisal model could also provide information on attributes affecting value in the neighbourhood.

With the government making moves to change the house selling process and place emphasis upon the vendor to source surveys and market valuations, the requirement for an independent judgement of market value as opposed to a mortgage risk approval process will become more necessary. With recent research showing that 90% of professional valuations are within 5% of the transaction price (Gronow, et al,

1996) - a figure clearly available at the time of valuation - then moves towards the derivation of true market value using a mathematical model will no doubt be welcomed.

8.9 Factors that Influence the Value of Residential Properties

The aim of this research was not to identify an exhaustive list of factors that have influence of the value of a residential property. However, it is worth summarising those factors that appeared to have the greatest influence on value investigated in this research.

From the Census data analysis, using both Kohonen and Genetic Algorithm methodologies, the following factors improved modelling accuracy:

- House Type
- Employment/Profession
- Tenure
- Car Availability
- Education

In addition, in one or two models, available amenities and ethnic makeup of the surrounding area helped to define homogeneous subsets. However, this influence appeared to be rare and can be explained by the monolithic nature of these measures within the geography investigated - leading to single clusters that have no bearing on value.

For the mixed Census analysis performed using the Genetic Algorithm approach a fairly consistent improvement in accuracy was achieved based on Employment; Tenure; and, Car Availability. It could be argued that these factors are all surrogates for wealth and therefore concluding that the overbearing influence on property value is indeed the relative wealth of the surrounding neighbourhood.

One of the suggestions that is made in the future work section of this thesis is the application of the techniques developed during this research program to more dynamic and representative data pertaining to regional wealth.

8.10 Overall Conclusions

Bringing together the many strands of this research a number of conclusions based on empirical results are evident:

- A set of models, each dedicated to a certain narrow domain, can significantly outperform predictions made by a single more general model trained on all of the available training data.
- Demand side data, like Census data, enhances the quality of the model. The geographical size of the Census area selected affects its performance as a locational surrogate in an appraisal model. The advised aggregate to use to represent location is the Enumeration District.
- Many different features describe the characteristics of an area. These include obvious features such as 'housing stock in neighbouring region', but also include less obvious features such as 'number of cars per household'.
- Models created from the stratification technique can be used to predict property values in other areas that have similar Census characteristics.

Based on the conclusions drawn within this chapter, the following provides direct answers to the original aims set at the outset of the research programme:

Assess empirically the suitability of Artificial Neural Network models to assist a residential property valuer in day-to-day valuations:

Neural networks have proven able to elicit functions from transaction data describing previous mortgage applications to a suitable standard as to estimate within a defined bounding the value of a previously unseen property. However, the accuracy of this estimation degrades significantly, as training data becomes more heterogeneous with respect to location and property type. To counteract this, empirical research has led to findings that Census characteristics are able to provide a degree of representation of neighbourhood and hence increase the accuracy found within neural network models. Furthermore, prototype systems using other techniques from the field of artificial intelligence has shown that this Census data can be most effectively used as a means of stratifying the heterogeneous dataset into homogeneous subsets that are more susceptible to training a neural network

(or MRA). Out of the stratification models investigated the genetic algorithm approach is most useful as it requires little expert valuation knowledge, is able to automatically improve its solution over time, and avoids generation of redundant clusters given its directional nature.

Use existing or develop new techniques that facilitate cognitive understanding of the underlying reasoning processes of ANN models:

This is one of the major disadvantages cited for using neural networks as modelling techniques within residential property literature. The issue being the difficulty in representing highly distributed numerical representations of data in a format suitable for human interpretation. Some work has been done in this field - however at the moment this is restricted to classification problems, as no useful methods are currently available for regression problems such as the estimation of property value. To provide the generated models with some aspect of human-readability, it was decided that time could profitably used by investigating techniques for eliciting rules from the stratification process as opposed to the neural network end nodes. Substantial progress was made in this respect with rules elicited both from the Kohonen approach using quartile descriptions and from the genetic algorithm approach via a decoding algorithm. It is also worth noting that the stratification approaches investigated provide homogeneous training data suitable for most regression models retaining the option to use a preferred method or interchange methods as appropriate.

Specify a prototype system that could be bolted onto a comparables database to provide ANN estimates of value:

This is by far the least significant aim of the research, and hence has only been afforded a very small amount of attention. Essentially, the research has shown that there is potential in neural networks and AI techniques for building residential property appraisal models that could run alongside more traditional analysis to provide professional valuers with additional evidence. Implementation could take the form of embedded code that mimics the representations found by a trained neural network, or perhaps a database component that could sit next to a transaction database and maintain a host of valuation functions by continually relearning using the most up-to-date

transaction data. The actual implementation is a decision that will be made at the appropriate time. However, it is envisaged that the concepts discussed in this thesis could run alongside the existing comparable approach providing the appraisal professional with the familiarity and dependency of his existing technique and the insight and stimulus provided by data analysis techniques. The benefit being that the data analysis techniques are able to uncover important patterns and trends that may otherwise be missed.

8.11 Contribution to Knowledge

The scope of this research encompasses the field of residential property appraisal and also the field of artificial intelligence. The research detailed in this thesis has made the following contributions to knowledge in these two fields:

Appraisal

Firstly, the use of tree induction techniques as a method of mass appraisal for property taxation was demonstrated. The advantage shown over other methods was the ease at which expert type rules could be extracted. Secondly, a high level definition of the stratification process by property type and location was defined using predicate logic. This enables a clear understanding of the theoretical problem to be gained. Thirdly, the value of the inclusion of additional data - such as Census data - into the appraisal model was demonstrated. Clearly the addition of more descriptive data can provide a more detailed dataset on which to build an appraisal model. Finally, an investigation of technical solution to automating the stratification processes led to the development of two methodologies that showed improvements in accuracy over the non-stratified model and also facilitated the extraction of rules describing some of the underlying functionality of the models.

Neural Networks and Artificial Intelligence

Within the field of artificial intelligence and specifically that of neural networks, this research has led to the development of methodologies for modelling data containing multiple functions. Specifically, the development of a methodology for including the Gamma Test within a genetic algorithm for dynamically stratifying a large data-set into sub-sets more suitable for use as training sets for neural networks could have application in other fields. The move towards amalgamating techniques into hybrid

intelligent models that as a whole compensate for the weakness in the individual technique (Goonatilake and Khebbal, 1995) has been demonstrated in this research. Models produced using the stratification methods proposed in this research can encompass genetic algorithms (stratification technique), neural networks (stratification and regression modelling), other regression techniques (subsitute MRA for MLPs in modelling of homogeneous sub-regions) and expert systems (rule extraction).

8.12 Suggestions for Future Work

What has become apparent throughout this research program is that there is an enormous scope for further work in the field of residential (and commercial) property appraisal. Worryingly, there appear to be more questions to answer now at the end of the research than ever seemed possible at the outset. The following is a list of new objectives that have arisen that the author would hope to investigate in the foreseeable future:

One of the major conflicts that have arisen of the past few years has been which technique to use as to build the regression model. Having developed a useful methodology for preparing training data suitable for generating accurate neural network models, it would be highly beneficial to begin a further research program to carry out such an investigation. Further research will benefit from having 'good' training sets readily available and focus can then be placed upon selecting the regression or classification model technique that provides the appraisal professional with the most information and the highest level of accuracy. (See Lam, 1996 for background information on a number of potential candidate regression modelling techniques that may form the basis of a future research programme).

The Census data used throughout the major empirical work provided a useful source of geodemographic information. However, it is quite legitimate to criticise this source of data for real-time valuations given the degradation that must occur in such periodical data. The Census data was of course used due to its ease of availability and direct join with property data via postcode. However, it would be beneficial to identify more dynamic data such as employment statistics, regional statistics, crime rates etc and use this together with subjective opinions formed by professional valuers as inputs to the stratification algorithm.

In setting the objectives for future research, focus is placed upon the use of dynamic 'demand side' data, utilising the most appropriate regression methodologies (e.g. MRA or ANNs). Hopefully, this focus combined with the stratification methodologies presented in this thesis will provide sufficient tools to begin the development of a commercial residential property appraisal system.

8.13 Final Remarks

The real test for any system claiming expertise equivalent to that of its human counterpart is to pitch one against the other. One day in the near future when the prototype techniques are sufficiently developed then it would be interesting to invite professional valuers to compare their estimates of value (without giving them transaction price as a starter) with those generated by an intelligent computer model.

9. REFERENCES

Adair, A S and McGreal, W S, 1986, The Direct Comparison Method of Valuation and Statistical Variability, Journal of Valuation, Vol. 5(1), pp 41-48.

Adair, A S, and McGreal, S 1987, The Application of Multiple Regression Analysis in Property Valuation, Journal of Valuation, Vol. 6, pp 57-67.

Adair, A S, Berry, J N, and McGreal, W S, 1996, Hedonic Modelling, Housing Sub-markets and Residential Valuation, Journal of Property Research, Vol. 13, pp 67-83.

Adair, A, and McGreal, S, 1995, Investigation of the Influence of Property and Socio-Economic Variables on residential Values and the Formulation of Valuation Models Based on Regression Analysis, Technical Report, Real Estate Studies Unit, School of the Built Environment, University of Ulster (April 1995).

Allen, WC and Zumwalt, JK, 1994, Neural Networks: A Word of Caution, unpublished working paper, Colorado State University.

Almond, NI, Lewis, OM, Jenkins, DH, Gronow, SA, and Ware, JA, 1997, Intelligent Systems for the Valuation of Residential Property, RICS Cutting Edge 97.

Almond, NI, 1999, The development of an holistic methodology for the valuation of residential property, Unpublished Ph.D. thesis - University of Glamorgan.

Almy, R, Horbas, J, Cusack, M, and Gloudemans, R, 1998, The Valuation of Residential Property using Regression Analysis, Computer Assisted Mass Appraisal: An International Review, Ed. McCluskey, WJ and Adair, AS, Ashgate Publishing Company, England.

Andrews, R, Cable, R, Diederich, J, Geva, S, Golea, M, Hayward, R, Ho-Stuart, C, and Tickle, A B, 1995, "An Evaluation and Comparison of Techniques for Extracting and Refining Rules from Artificial Neural Networks, in Knowledge-Based Systems Journal Vol 8, No 6 (December 1995).

Antwi, A, 1995, Hedonic Price Indices, Estates Gazette, Vol. 9543, pp 124-126.

Ashton, P, 1972, The Use of Multiple Regression Analysis in the Valuation of Real Estate, The Real Estate Appraiser, January 1972, pp 12-14.

Bellman, R, 1961, Adaptive Control Processes: A Guided Tour. New Jersey: Princetown University Press.

Bigus, JP, 1996, Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support, McGraw-Hill

Bishop, CM, 1994, Neural Networks for Pattern Recognition, Oxford University Press. ISBN 0-19-853864-2.

Borst, R A, 1991, Artificial Neural Networks: The Next Modelling / Calibration Technology for the Assessment community?, Property Tax Journal, Vol.10, pp 69-94.

Borst, R A, 1993, A Method for the Valuation of Residential Properties using Artificial Neural Networks in Conjunction with Geographical Information Systems.

Bourassa, SC, Hamelink F, Hoesli, M, and MacGregor, BD, 1997, Defining Residential Sub-Markets: Evidence from Sydney and Melbourne, Real Estate Research Unit, Department of Property, Faculty of Architecture, Property and Planning, University of Auckland, Private Bag 92019, Auckland 1, New Zealand, ISBN 1-877183-02-4

Boyle, A, 1982, An Expert System of Valuation of Residential Properties, 2 JV 3, 271-286.

Bruce, R W and Sundell, D J, 1977, Multiple Regression Analysis: History and Applications in the Appraisal Profession, The Society of Real Estate Appraisers, Vol. 43, pp 37-44.

CACI, 1989, "ACORN Classifications", CACI Limited, CACI House, Kensington Village, Avonmore Road, London. W14 8TS

Carbone, R, and Longini, R, L, 1977, A Feedback Model for Automated Real Estate Assessment, Management Science, Vol. 24, pp 214-248.

Charlton M, Openshaw S, and Wymer C, 1985, Some New Classifications of Census Enumeration Districts in Britain: A Poor Man's ACORN, Journal of Economic and Social Measurement, Vol 13, pp 69-98.

Chen, Ke, Xiang, Yu, Huisheng, and Chi, 1997, Combining Linear Discriminant Functions with Neural Networks for Supervised Learning, Journal of Neural Computing and Applications, Vol. 6, Springer-Verlag, London.

Clancy W J, 1983, The Epistemology of a Rule-based Expert System: A Framework for Explanation, Artificial Intelligence, Vol 20, pp 215-251.

Collins, A, and Evans, A, 1994, Aircraft Noise and Residential Property Values: an Artificial Neural Network Approach, Journal of Transport Economics and Policy, pp 175-197

Colwell, PF and Foley, KW, 1979, Electricity Transmission Lines and the Selling Price of Residential Property, *The Appraisal Journal*, Vol. 37, pp 283-288.

Craven, MW, 1996, Extracting Comprehensible Models from Trained Neural Networks, Unpublished Ph.D. Thesis, University of Wisconsin-Madison

Czernkowski, R, 1990, Expert Systems in Real Estate Valuation, Journal of Valuation, Vol. 8(4), pp 376-393.

Dale A, and Marsh, C, 1993, The 1991 Census User's Guide, HMSO Publications

Davis R, Buchanan B, and Shortcliffe E, 1977, Production Rules as a Representation for a Knowledge - Based Consultation Program, Artificial Intelligence, Vol 8, No 1, pp 15-45.

Diederich J, 1989, Explanation and Artificial Neural Networks, German National Research Centre for Computer Science (GMD), W.Germany.

Do, Q, and Grudnitski, G, 1992, A Neural Network Approach to Residential Property Appraisal, The Real Estate Appraiser, December 1992, pp 38-45.

Dodgson, J and Topham, N, 1990, Valuing Residential Properties with the Hedonic Method: A Comparison with the results of Professional Valuations, Housing Studies 5 (3), pp 209-213.

Dodgson, J, 1989, Valuing Residential Properties with the Hedonic Method: A Comparison with the Results of Professional Valuations, Occasional Paper - University of Salford.

DTI, 1990, DTI Guidelines for Neural Computing, DTI's Neural Computing Technology Transfer (NCTT) Programme.

Eberhart, R C and Dobbins, R W, 1990, Case Study I: Detection of Electroencephalogram Spikes in Neural Network PC Tools, California: Academic Press, Inc.

Eckert, JK, 1990, Property Appraisal and Assessment Administration, The International Association of Assessing Officers, Chicago, I11,

Evans, A, James, H, and Collins, A, 1992, Artificial Neural Networks: an Application to Residential Valuation in the UK, Journal of Property Valuation and Investment, Vol. 11, pp 195-204.

Fahlman, S and Lebiere, C, 1988, Faster Learning Variations on Back-Propagation: An Empirical Study. In Dtouretzky, G E Hinton and T J Sejnowski (Eds), San Mateo, CA: Morgan Kaufmann.

Flemming M.C.and Nellis J.G. (1994) The Halifax Price Index technical details, Halifax Building Society.

Garson, GD, 1991, Interpreting Neural Network Connection Weights, AI Expert, Vol.6, pp 47-51.

Goldberg, DE, 1989, Genetic Algorithms in Search Optimisation and Machine Learning, Addison Wesley

Goonatilake S, and Khebbal S, 1995, Intelligent Hybrid Systems, John Wiley and Sons.

Goulden, CH, 1989, Methods of Statistical Analysis, John Wiley and Sons

Grant, CA, and McTear, MF, 1992, An Expert System for Property Valuation, Journal of Property Valuation and Investment, Vol. 10

Greaves, M, 1984, The Determinants of Residential Values: The Hierarchical and Statistical Approaches, 3 JV, pp 5-23

Gronow SA, Ware JA, Jenkins DH, Lewis OM, and Almond NI, 1996, "A Comparative Study of Residential Valuation Techniques and the Development of a House Value Model and Estimation System", ESRC ROPA Report.

Gronow, S and Scott, I, 1985, Expert Systems, Estates Gazette, Vol. 276, pp. 1012-1014.

Gronow, S and Scott, I, 1986, Expert systems and Multiple Regression Analysis, Estates Gazette, Vol. 278, pp. 694-695.

Gronow, S and Scott, I, 1987, Expert Systems - Knowledge Representation for Building Society Mortgage Valuations, Journal of Valuation, Vol. 6(1), pp. 87-101.

Guntermann, KL and Colwell, PF, 1983, Property Values and Accessibility to Primary Schools, *The Appraisal Journal*, Vol. 49, pp 62-68.

Have, GM ten, Veld, AG op't, and Janssen, JE, 1998, TAXES: Residential Property Valuation for Local Tax Purposes in the Netherlands, Computer Assisted Mass Appraisal: An International Review,

Hecht-Nielson, 1990, Neurocomputing, Reading MA, Addison-Wesley.

Holland J, 1987, Genetic Algorithms and Classifier Systems: Foundations and Future Directions, In Genetic Algorithms and their Applications: Proceedings of the 2nd International Conference on Genetic Algorithms.

Jackson P, 1986, Introduction to Expert Systems, Addison Wesley.

James, H, 1994, An 'Automatic Pilot' for Surveyors, RICS Cutting Edge.

Jenkins DH, 1992, Expert Systems in the Land Strategy of Cardiff City Council, Unpublished MPhil. University of Glamorgan.

Jones, AJ 1996, The Gamma Test, Department of Computer Science, University of Wales, Cardiff, UK.

Kidd A, 1986, Knowledge Acquisition for Expert Systems: A Practical Handbook, Plenum Press.

Kohonen, T, (1984), A Simple Paradigm for the Self-Organised Formation of Structured Feature Maps, in Competition and Co-operation in Neural Networks. ed. S. Amari, M. Arbib. vol. 45. Berlin: Springer Verlag.

Lam, ETK, 1996, Modern Regression Models and Neural Networks for Residential Property Valuation, RICS Cutting Edge

Lawrence, J, 1992, Introduction to Neural Networks, California Scientific Press, Nevada City, California, USA.

Lenk, MM, Worzala, EM, and Silva, A, 1997, High-Tech Valuation: Should Artificial Neural Networks Bypass the Human Valuer, Journal of Property Valuation and Investment, Vol. 15, No.1, pp 8-26.

Lessinger, J, 1969, Econometrics and Appraisal, Appraisal Journal, Vol. 37, pp 501-512.

Li, MM and Brown, HJ, 1980, Micro-Neighbourhood Externalities and Hedonic House Prices, Land Economics, 56, pp 125-140.

Lu, M T, and Lu, D H, 1992, Neurocomputing Approach to Residential Property Valuation, Journal of Microcomputer Systems Management, Vol. 4(2), pp 21-30.

Mackmin, D, 1994, The Valuation and Sale of Residential Property, Routledge, 2nd Edition.

Matysiak, GA, 1991, Comment on: Valuation by Comparable Sales and Linear Algebra, Journal of Property Research, Vol. 8, pp 21-27.

Matysiak, GA, 1992, Econometrics, Linear Programming and Valuation: Reply, Journal of Property Research, Vol. 9, pp 114-121.

McCluskey, WM, 1996, Predictive Accuracy of Machine Learning Models for Mass Appraisal of Residential Property, New Zealand Valuers' Journal, (July 1996) pp 41-47

McClusky, WJ and Adair, AS, 1998, Computer Assisted Mass Appraisal: An International Review, Ashgate Publishing Company, England.

McGreal, S, Adair, A, McBurney, A, and Patterson, D, 1998, Neural Networks: the Prediction of Residential Values, Journal of Property Valuation and Investment, Vol. 16, No.1

Michalski R, and Chilansky R L, 1980, Learning by Being Told and Learning from Examples, Journal of Policy Analysis & Information Systems, Vol 4.

Miller, N G, 1982, Residential Property Hedonic Pricing Models: A Review, Research in Real Estate, Vol. 2, pp 31-56.

Millington, A F, 1994, An Introduction to Property Valuation, Estates Gazette, 4th Edition.

Mozer, M C, and Smolensky, P, 1989, "Using Relevance to Reduce Network Size Automatically", Connection Science, 1, 3-16.

Munro, M, 1986, Testing for Segmentation in the Glasgow Private Housing Market, Discussion Paper No 8, Centre for Housing Research, University of Glasgow, Glasgow.

Nawawi, AH, Jenkins, DH, and Gronow, SA, 1996, Computer Assisted Rating Valuation of Commercial and Industrial Properties in Malaysia: Developing an Expert System from a Multiple Experts Knowledge Elicitation Methodology, RICS, Cutting Edge

Newell, GJ, 1982, The Application of Ridge Regression to Real Estate Appraisal, The Appraisal Journal, Jan 1992, pp 116-119

Openshaw, S and Wymer, C, 1994, Classification and Regionalisation, in S Openshaw (ed), Census User's Handbook, Longmans, London.

Pennington, G, Topham, N, and Ward, R, 1990, Aircraft Noise and Residential Property Values Adjacent to Manchester International Airport, Journal of Transport Economics and Policy, 25(1), pp 49-59.

Quinlan J R, 1986, Induction of Decision Trees, Machine Learning, Vol 1, pp 81-106.

Quinlan, J R, 1993, C4.5: Programs for Machine Learning. Morgan Kaufmann 1993, ISBN 1-55860-283-0

Rayburn, WB and Tosh, DS, 1995, Artificial Intelligence: The Future of Appraising, The Appraisal Journal (October 1995).

Refenes, AN, 1994, Neural Networks in the Capital Markets, Wiley and Sons, Chichester, England.

Rumelhart, DE and McClelland, JL, 1986, Parallel Distributed Computing, Ch.5 MIT Press, Cambridge Mass.

Sauter, B,1985, Solving Today's Computer Assisted Valuation Issues Using the Adaptive Estimation Procedure and Bayesian Regression, paper presented at the Second World Congress on Computer Assisted Valuation, Lincoln Institute of Land Policy, Massachusetts.

Schoneburg, E, 1990, Stock Price Prediction Using Neural Networks: A Project Report, Neurocomputing 2, Elsevier Science Publishers, pp 17-27.
Schwartz, T J, 1995, Automating Appraisal, Wall Street and Technology, Vol. 12, No. 13, pp 64-66.

Schwartz, T J, 1995, Automating Appraisal, Wall Street and Technology, Vol. 12, No. 13, pp 64-66.

Shenkel, W, 1978, Modern Real Estate Appraisal, McGraw-Hill, New York and London.

Sleight, P, 1993, Targeting customers: how to use geodemographic and lifecycle data in your business, NTC Publications, Henley on Thames.

Southwick R, 1991, Explaining Reasoning: An Overview of Explanation in Knowledge-Based Systems, The Knowledge Engineering Review, Vol 6, No 1, pp 1-19.

Spivey, J.M., 1992, The Z Notation: a Reference Manual, Prentice Hall International, Second Edition.

Stefansson, A, Koncar, N, and Jones, AT, 1997, A Note on the Gamma Test, Journal of Neural Computing and Applications, Vol.5 No. 3, Springer Verlag

Strazheim, MR, 1973, Estimation of the Demand for Urban Housing Services from Interview Data, Review of Economics and Statistics 55 pp 1-8

Tay, D P H, and Ho, D K H, 1992, Intelligent Mass Appraisal, Journal of Property Tax Assessment and Administration, Vol. 10, pp 5-25.

Tay, DPH, and Ho, DKH, 1991, Artificial Intelligence and the Mass Appraisal of Residential Apartments, Journal of Property Valuation and Investment, 10;2, pp 525-540.

Tazelaar, J M, 1989, Neural Networks, BYTE, August p214.

Vermuri and Rogers 1994, Artificial Neural Networks Forecasting Time Series, IEEE Computer Society Press, California

Waggert, S, 1997, Gamma Test Documentation available from Prof. A. Jones at the Department of Computer Science, University of Wales, Cardiff, UK.

Weigand, AS, Rumelhart, D, and Huberman, B, 1991, Generalisation by Weight Elimination Applied to Currency Exchange Rate Prediction, Proc. IJCNN'91 IEEE press.

Werbos, p, 1974, Beyond Regression: New Tools for Prediction Analysis in Behavioural Sciences, Unpublished PhD. Thesis, Harvard University, Cambridge MA.

Wiltshaw, DG, 1991 a, Valuation by Comparable Sales and Linear Algebra, Journal of Property Research, Vol. 8, pp 3-19.

Wiltshaw, DG, 1991 b, Econometrics, Linear Programming and Valuation, Journal of Property Research, Vol.8 pp 123-132.

Wiltshaw, DG, 1993, Imperfect Price Information and Valuation by Comparable Sales, Journal of Property Research, Vol.10 pp 85-96.

Worzala, E, Margarita, L, and Silva, A, 1995, An Exploration of Neural Networks and Its Application to Real Estate Valuation, The Journal of Real Estate Research, pp 185-201.

Zirilli, JS, 1997, Financial Prediction Using Neural Networks, John Wiley and Sons, ISBN 1-850-32234-1

Zurada, J M, 1992, Introduction to Artificial Neural Systems, West Publishing Company (ISBN 0-314-93391-3) p58

APPENDIX 1 - DATA SCHEMAS

A1.1 Schema of Mortgage Transaction Database

Name of Field	Range or Example Value	Redundant Field	Used in ANN Models
Street Name	Newport Road	No	
District or Village	Roath	No	
City	Cardiff	No	
Comparable ID	Numeric link to database	No	
Street ID	Numeric link to street dB	No	
Postcode	CF37 4HG	No	
Source	Valuers Name	No	
Firm	Halifax/ Colleys	No	
Status	Blank Field	Yes	
New Build	True/ False	No	
Unit	1 - 6	No	X
Unit Type	Mid terraced, etc.	No	
Unit Size	Area M^2	No	X
Valuation Date	19 May 1995	No	
Main Heating	Gas, Elec. Etc.	No	X
Purchase Price	£45,000	No	
Number of Bedrooms	1 - 8	No	X
Number of Living Rooms	Blank Field	Yes	
Condition Risk Factor	1 - 5	No	
Quality of Evidence	Blank Field	Yes	
Sought After	True/ False	No	
Unit Name	Text String	No	
Unit Number	Numeric	No	
Plot Number	Blank Field	Yes	
Transaction ID	-1	Yes	
Age in Years	0 - 500	No	X
Extent	Blank Field	Yes	
Number of Garages	0 - 2	No	X
Quality	Average	Yes	
Construction Description	Blank Field	Yes	
Single Overriding Factor	Blank Field	Yes	
Local Authority	True/ False	No	
Traditional	True/ False	No	
Orientation	Blank Field	Yes	
Builder	Blank Field	Yes	
Tenure	1 - 4	No	X
Transaction Type	Open Market	Yes	
Value	0 - 255,000	No	X - dependent variable
Unit Style	Blank Field	Yes	
Number of Storeys in Unit	Blank Field	Yes	
Lowest Floor Unit	Blank Field	Yes	
Business Premises	False/ True	No	
Lease Term Unexpired	Numeric	No	
Rent	Currency	No	
Rent Indexation	Blank Field	Yes	
Service Charge	Currency	No	
Roll Number	Text Field	Yes	
Original Term	Numeric	No	
Remarks	Text Field (free form)	No	
Branch Code	0 - Blank Field	Yes	
Garage Spaces	Blank Field	Yes	

A1.2 Schema of Selected Census Database

The Census variables are listed within common groupings. Below the table there is a list of titles for the empirical work undertaken using the Census data in this Thesis together with a usage code indicating the attributes used in each empirical study. In the Table, the codes appear against the group title if all variables in that group were used and against individual attributes for empirical work not involving all variables from a particular Census grouping.

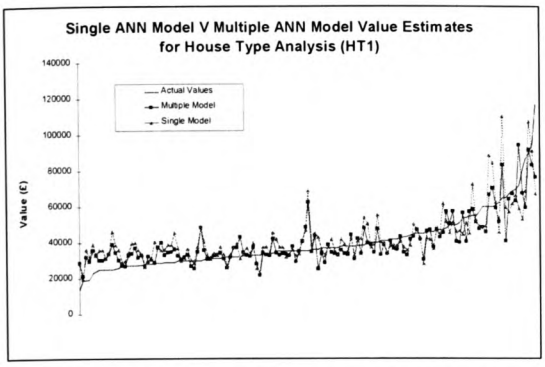
Socio-Economic Group	A,G		
Employers and Managers (Large est.)		Employers and Managers (small est.)	
Professional workers (self-employed)		Professional workers (employees)	
Ancillary workers and Artists		Foreman and Supervisors (non-manual)	
Junior non-manual workers		Personal Services workers	
Foreman and Supervisors (manual)		Skilled Manual workers	
Semi-Skilled Manual workers		Unskilled Manual workers	
Members of Armed Forces			
Employment	A,C		
full-time Employment	I,J	On Government Scheme	
part-time Employment		Unemployed	I,J
Self Employed			
Qualifications	A,H		
Qualified Persons		Higher Degree	
Degree	I,J	Diploma	
Qualified and on Government Scheme		Qualified and Unemployed	
Age Ranges of Qualified Persons			
Housing Stock	A,B,I, J		
Detached Properties		Purpose-Built Flats	
Semi-Detached Properties		Converted Flats	
Terraced Properties		Bedsits	
Tenure	A,D		
Owner Occupied (Outright)	I,J	Owner Occupied (Buying)	I,J
Privately Rented (Furnished)		Privately Rented (Unfurnished)	
Rented from Housing Association	I,J	Rented from Local Authority	I,J
Amenities	A		
Shared Use of WC		Exclusive Use of WC	
Central Heating			
Availability of a Car	A,E,I, J		
Households with no car		Households with 1 car	
Households with 2 cars		Households with 3+ cars	
Ethnicity	A,F		
White		Black Caribbean	
Black African		Black Other	
Indian		Pakistani	
Bangladeshi		Chinese	
Asian		Persons born in Ireland	
Miscellaneous Variables	A,		
Working Mothers (Part-Time)	I,J	Working Mothers (Full-Time)	
Lifestages (age ranges of residents)		Overcrowding (persons per household)	I,J
Travel to work estimates			

Key to Usage Coding

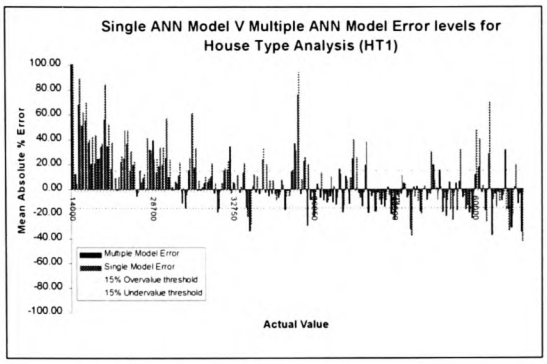
A - Enumeration District Analysis (Chapter 4)
B - House Type Analysis (Chapter 5 & 6)
C - Employment Analysis (Chapter 5 & 6)
D - Tenure Analysis (Chapter 5 & 6)
E - Car Availability Analysis (Chapter 5 & 6)
F - Ethnicity Analysis (Chapter 5 & 6)
G - Socio-Economic Analysis (Chapter 5 & 6)
H - Education Analysis (Chapter 5 & 6)
I - Two County Analysis (Chapter 5)
J - Mixed Census Analysis (Chapter 6)

APPENDIX 2 - GRAPHS SHOWING RESULTS OF THE KOHONEN STRATIFICATION METHOD

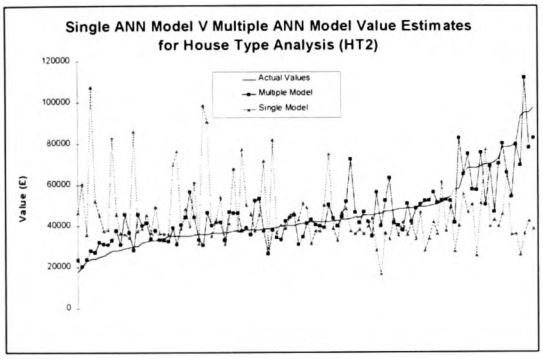
A2.1 House Type Analysis



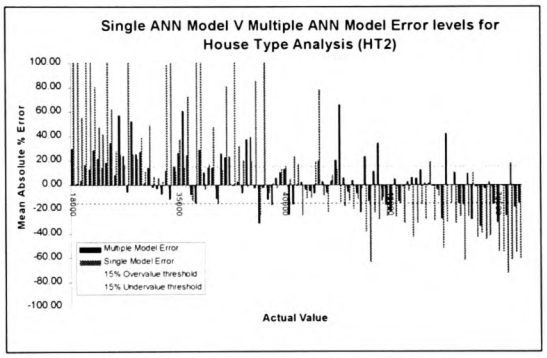
House Type Analysis HT1 (Predictions)



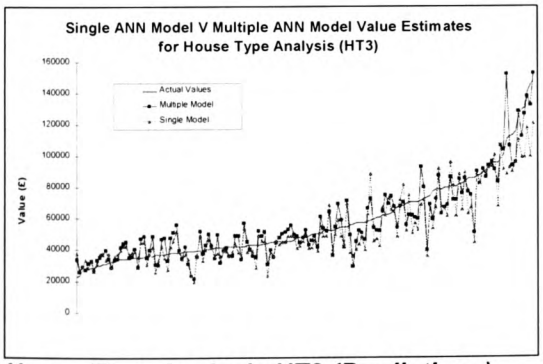
House Type Analysis HT1 (Errors)



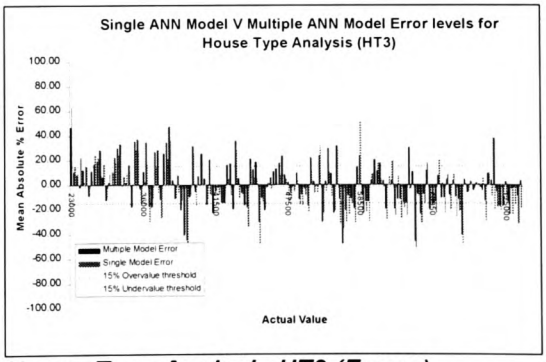
House Type Analysis HT2 (Predictions)



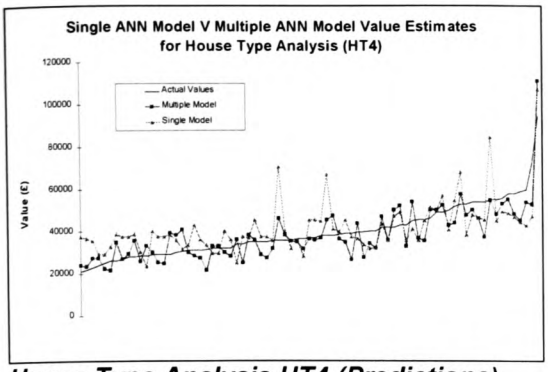
House Type Analysis HT2 (Errors)



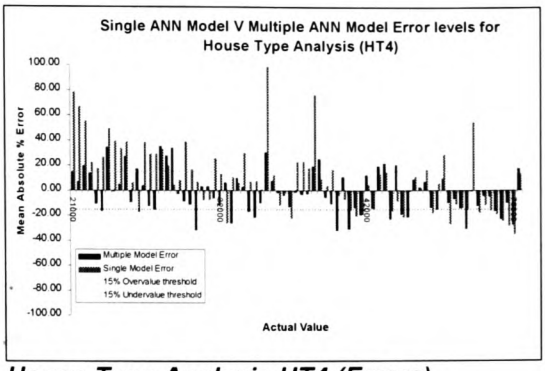
House Type Analysis HT3 (Predictions)



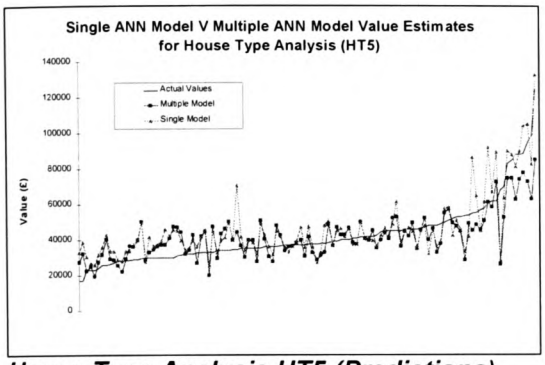
House Type Analysis HT3 (Errors)



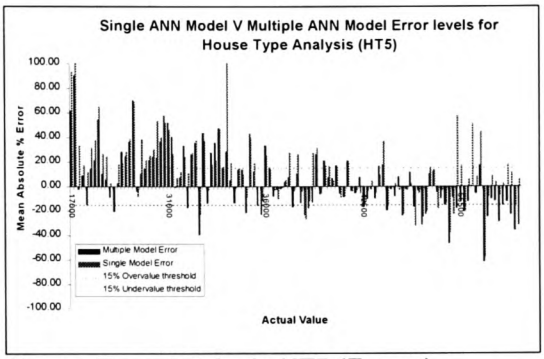
House Type Analysis HT4 (Predictions)



House Type Analysis HT4 (Errors)

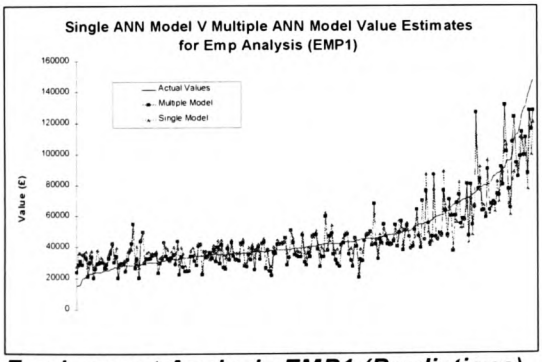


House Type Analysis HT5 (Predictions)

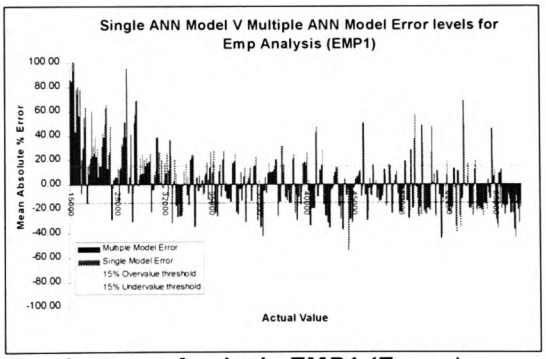


House Type Analysis HT5 (Errors)

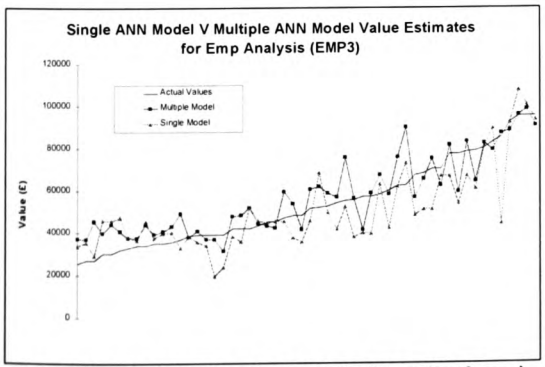
A2.2 Employment Analysis



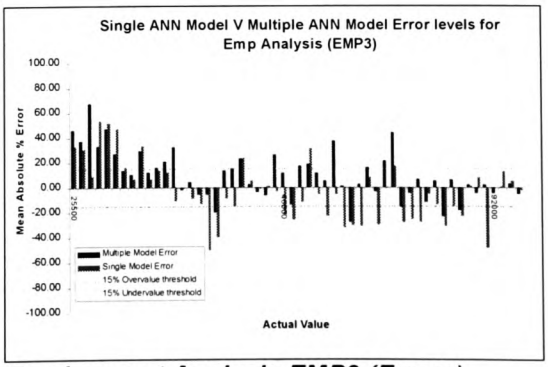
Employment Analysis EMP1 (Predictions)



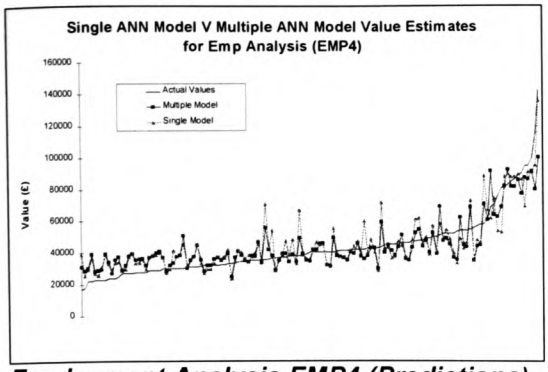
Employment Analysis EMP1 (Errors)



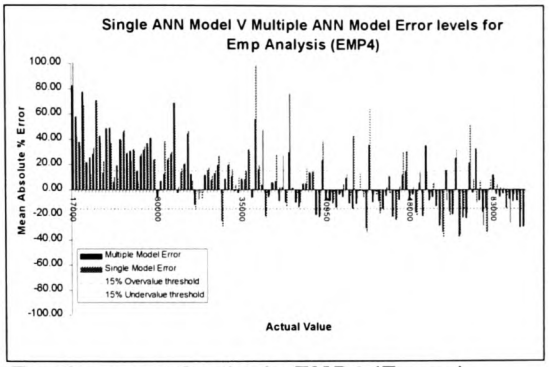
Employment Analysis EMP3 (Predictions)



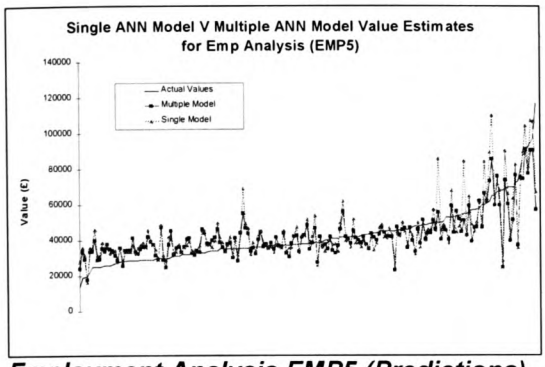
Employment Analysis EMP3 (Errors)



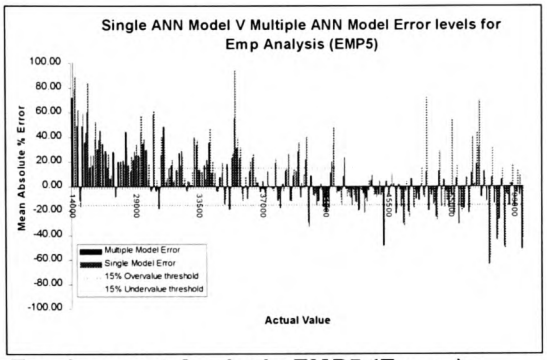
Employment Analysis EMP4 (Predictions)



Employment Analysis EMP4 (Errors)

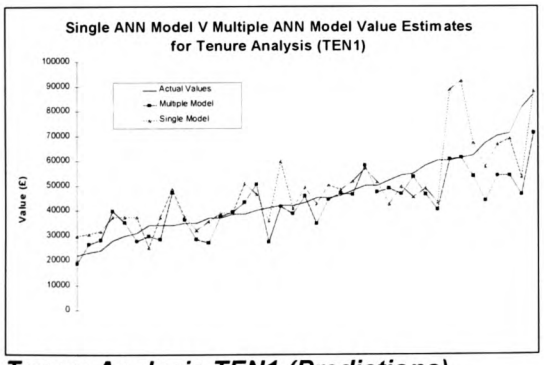


Employment Analysis EMP5 (Predictions)

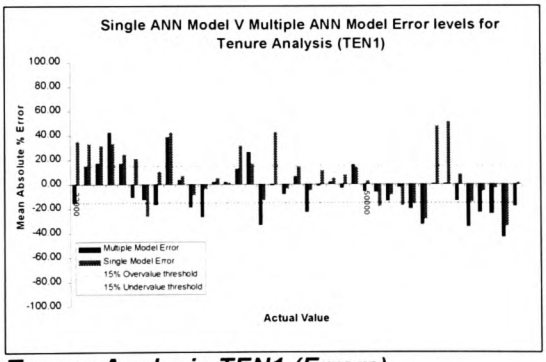


Employment Analysis EMP5 (Errors)

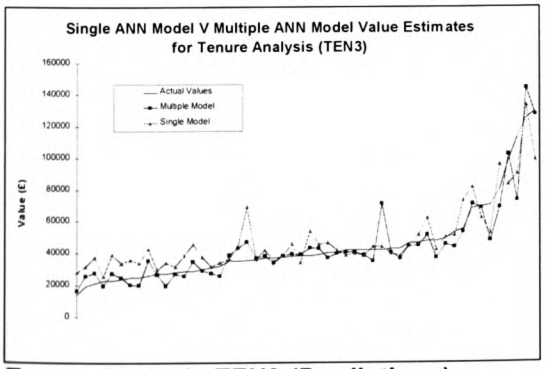
A2.3 Tenure Analysis



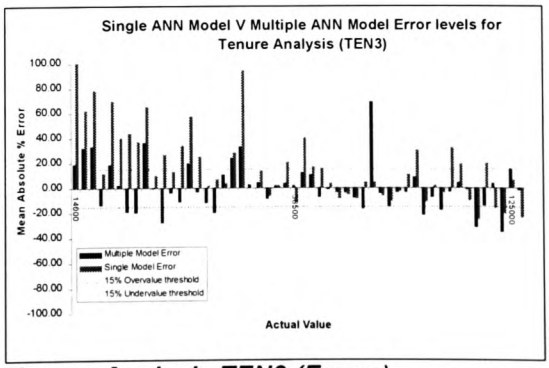
Tenure Analysis TEN1 (Predictions)



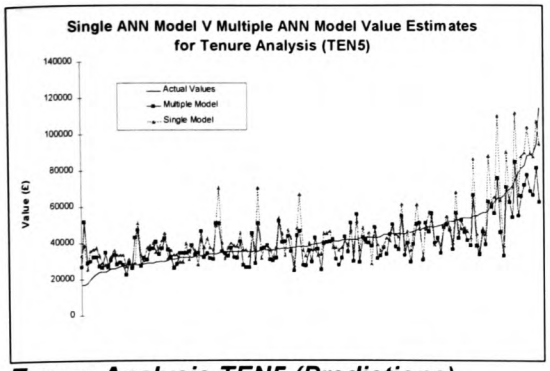
Tenure Analysis TEN1 (Errors)



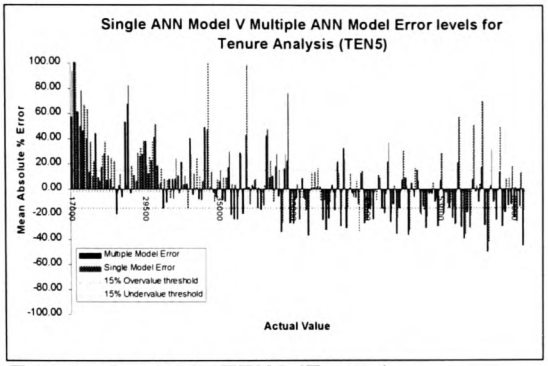
Tenure Analysis TEN3 (Predictions)



Tenure Analysis TEN3 (Errors)

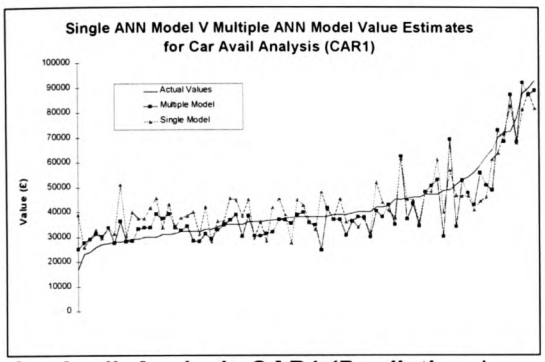


Tenure Analysis TEN5 (Predictions)

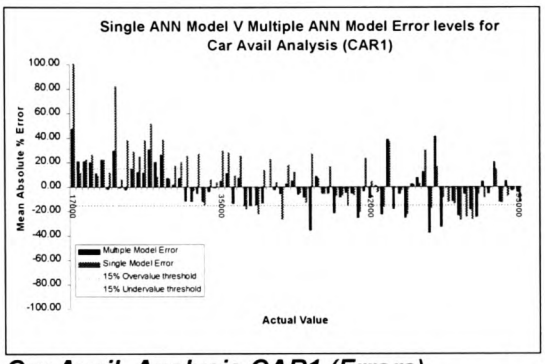


Tenure Analysis TEN5 (Errors)

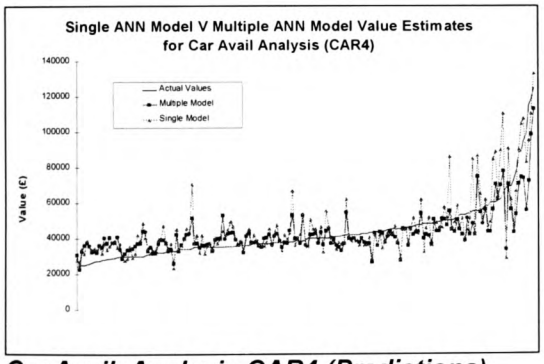
A2.4 Car Availability Analysis



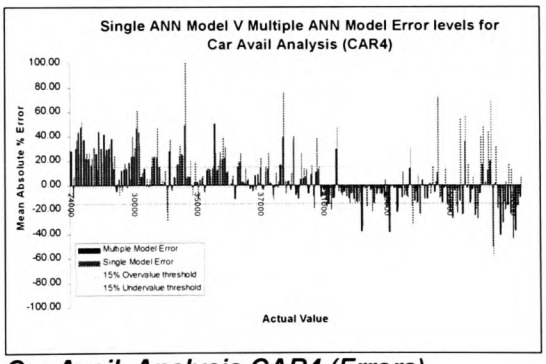
Car Avail. Analysis CAR1 (Predictions)



Car Avail. Analysis CAR1 (Errors)

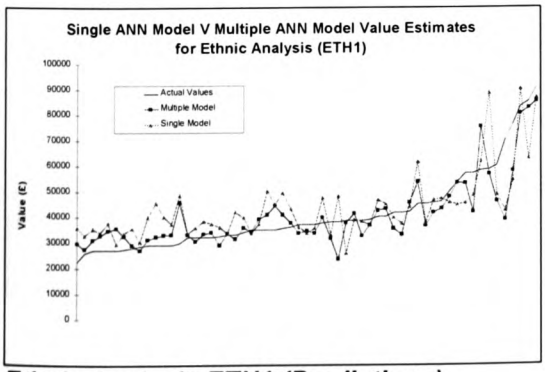


Car Avail. Analysis CAR4 (Predictions)

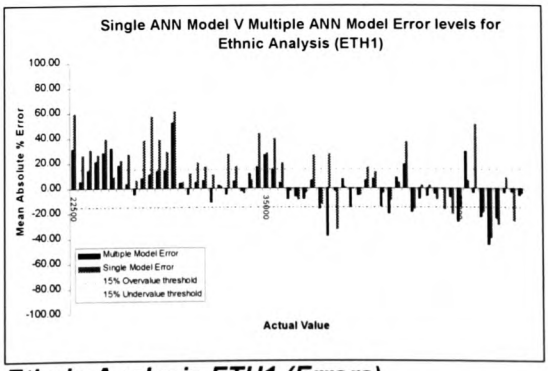


Car Avail. Analysis CAR4 (Errors)

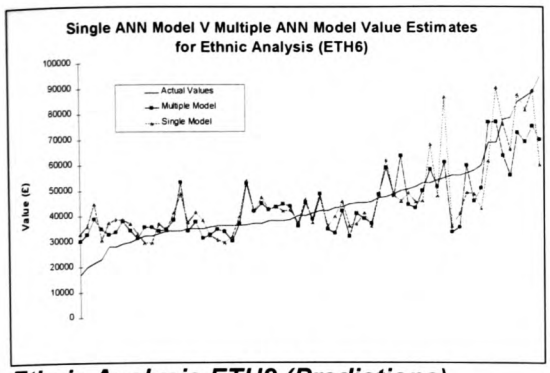
A2.5 Ethnic Analysis



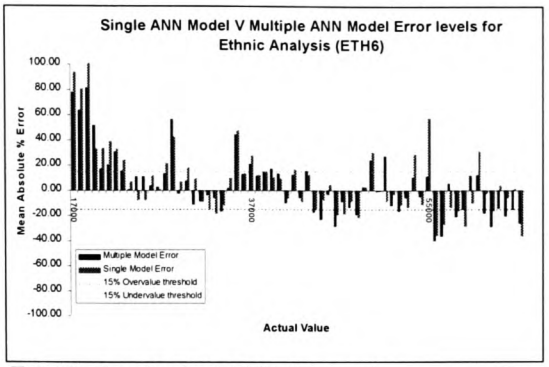
Ethnic Analysis ETH1 (Predictions)



Ethnic Analysis ETH1 (Errors)

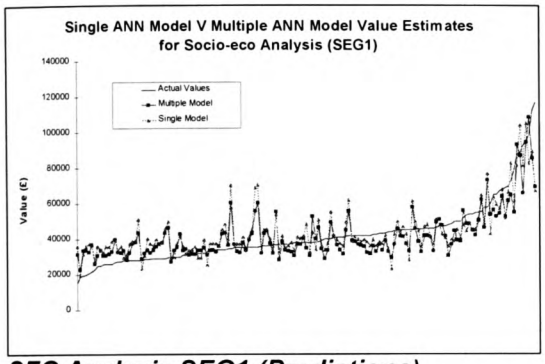


Ethnic Analysis ETH2 (Predictions)

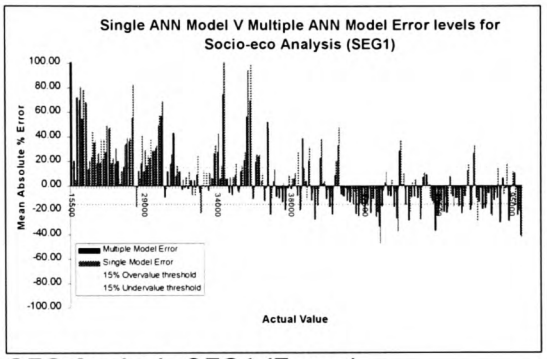


Ethnic Analysis ETH2 (Errors)

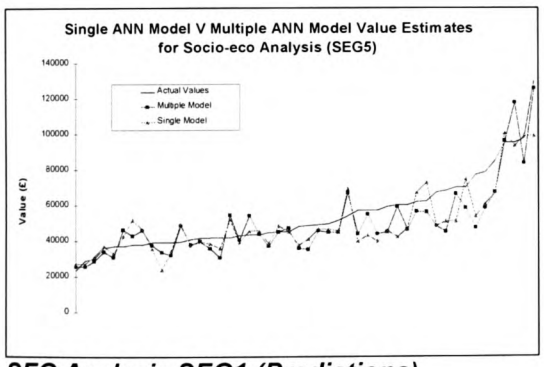
A2.6 Socio-Economic Analysis



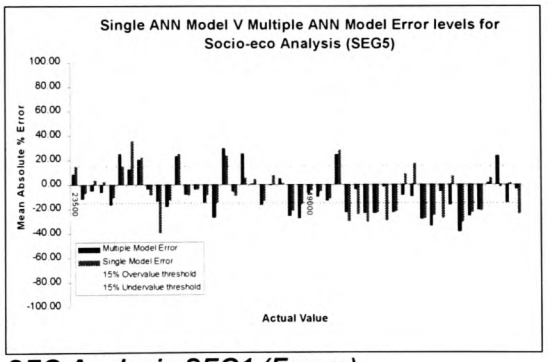
SEG Analysis SEG1 (Predictions)



SEG Analysis SEG1 (Errors)

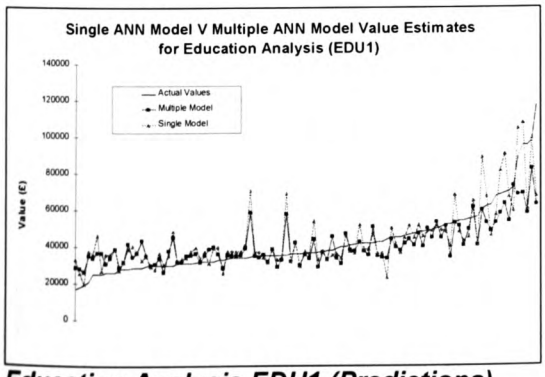


SEG Analysis SEG1 (Predictions)

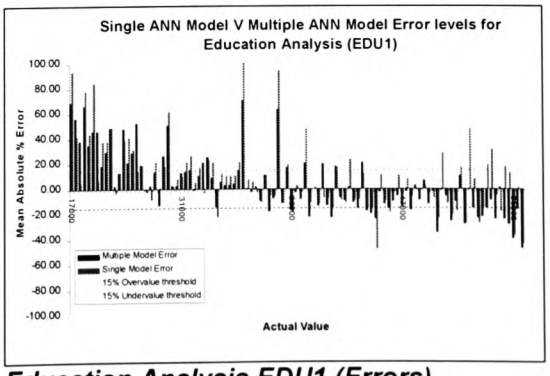


SEG Analysis SEG1 (Errors)

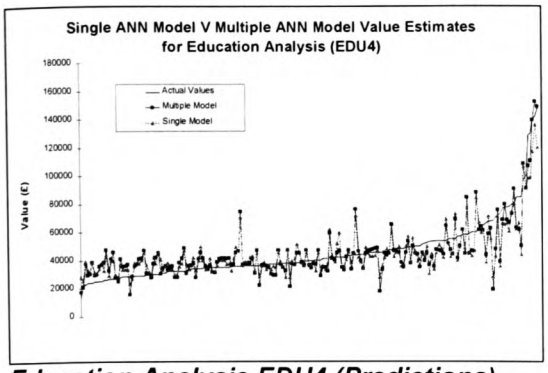
A2.7 Education Analysis



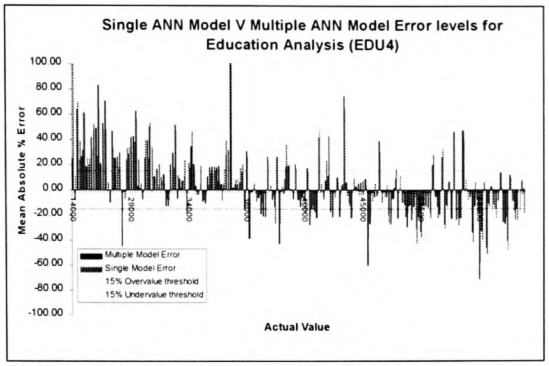
Education Analysis EDU1 (Predictions)



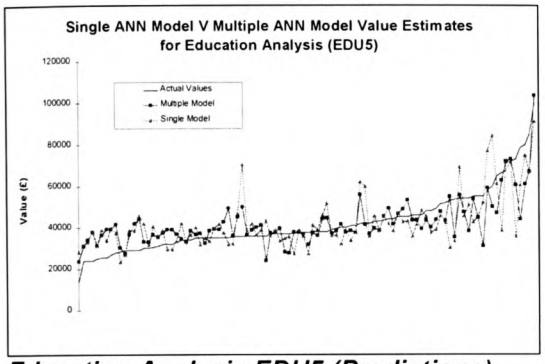
Education Analysis EDU1 (Errors)



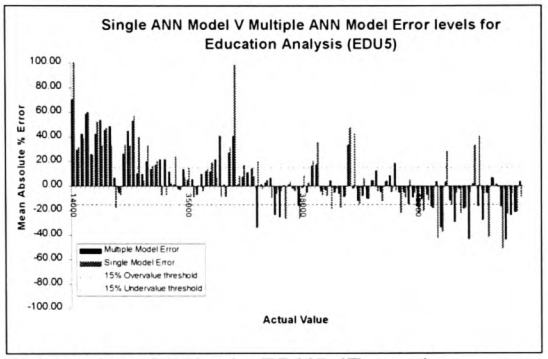
Education Analysis EDU4 (Predictions)



Education Analysis EDU4 (Errors)



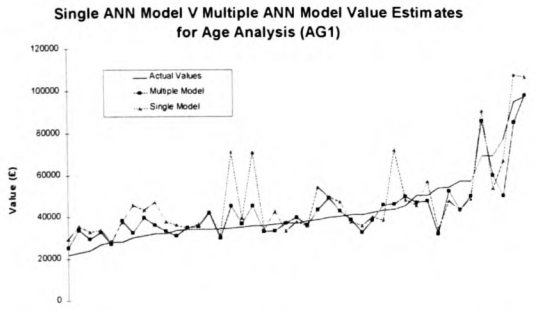
Education Analysis EDU5 (Predictions)



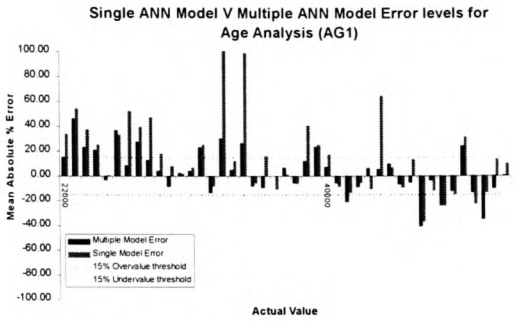
Education Analysis EDU5 (Errors)

APPENDIX 3 - GRAPHS SHOWING RESULTS OF THE GENETIC ALGORITHM STRATIFICATION METHOD

A3.1 Residents Age Analysis

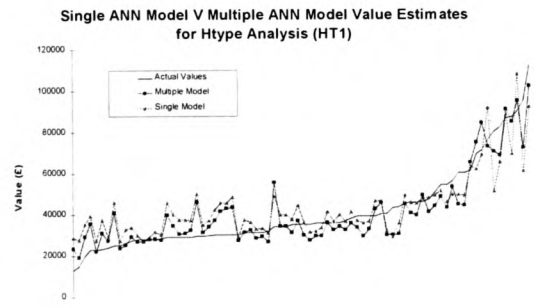


Res. Age Analysis AGE1 (Predictions)

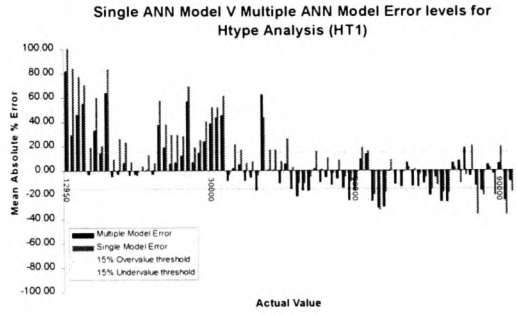


Res. Age Analysis AGE1 (Errors)

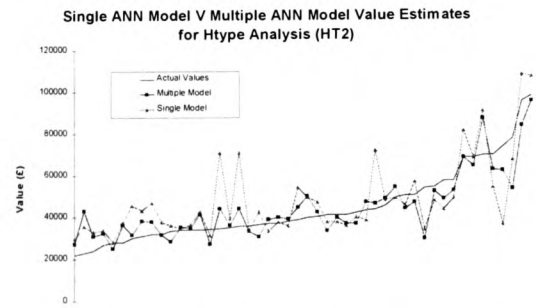
A3.2 House Type Analysis



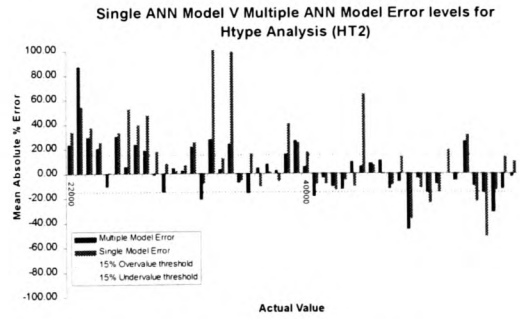
House Type Analysis HT1 (Predictions)



House Type Analysis HT1 (Errors)

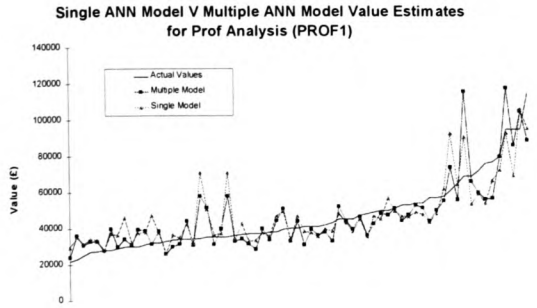


House Type Analysis HT2 (Predictions)

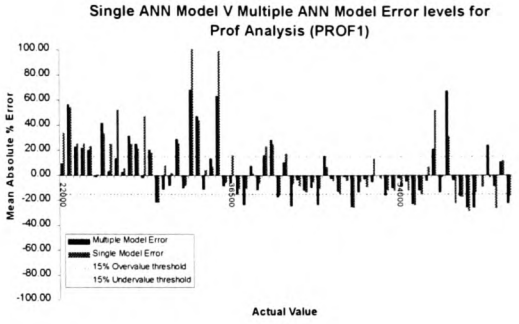


House Type Analysis HT2 (Errors)

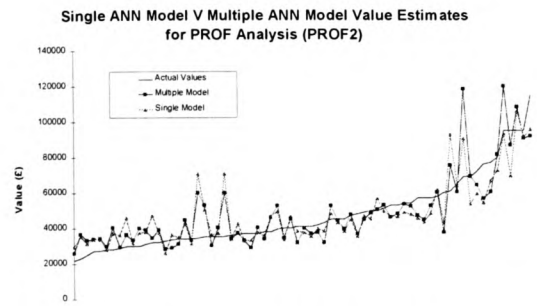
A3.3 Profession Analysis



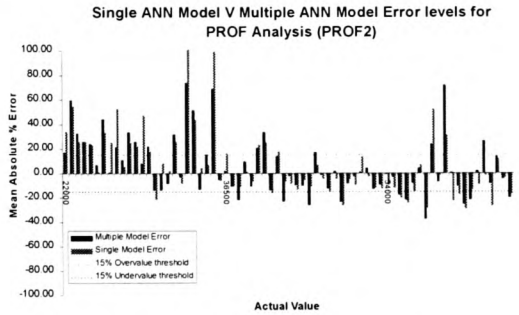
Profession Analysis PROF1 (Predictions)



Profession Analysis PROF1 (Errors)

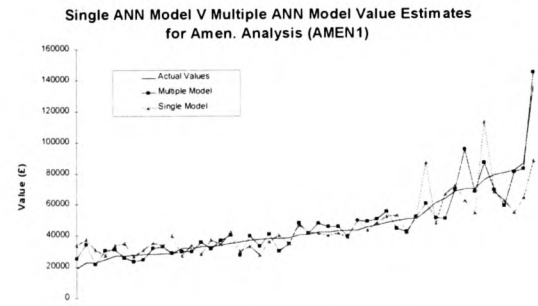


Profession Analysis PROF2 (Predictions)

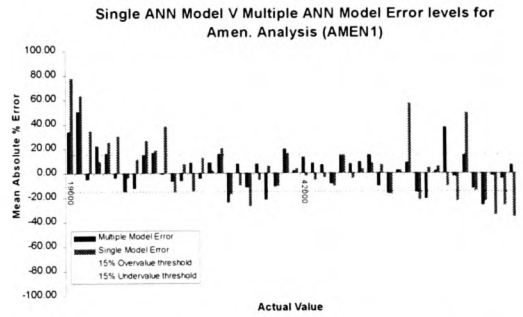


Profession Analysis PROF2 (Errors)

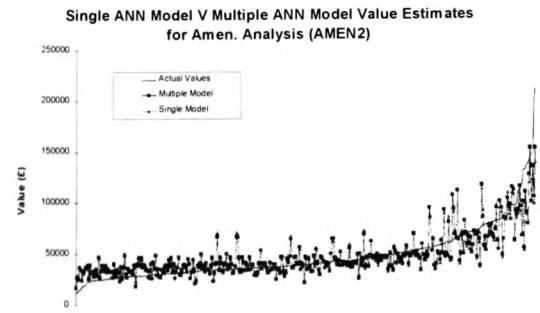
A3.4 Amenities Analysis



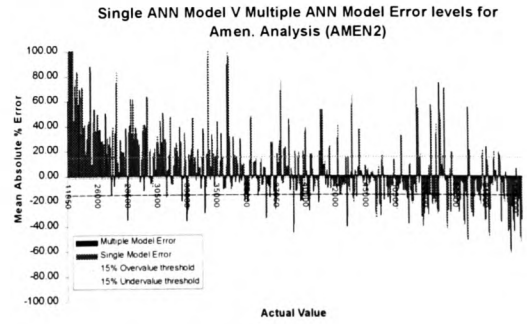
Amenities Analysis AMEN1 (Predictions)



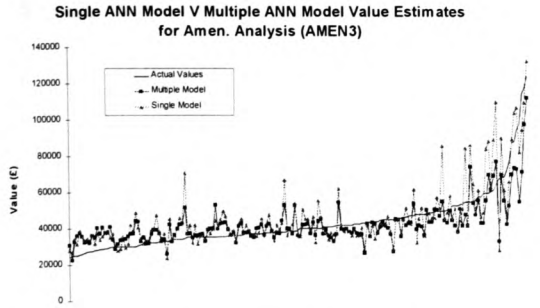
Amenities Analysis AMEN1 (Errors)

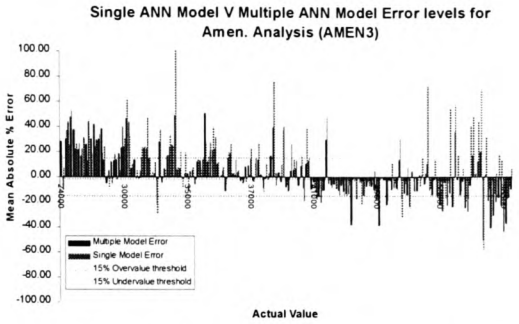


Amenities Analysis AMEN2 (Predictions)



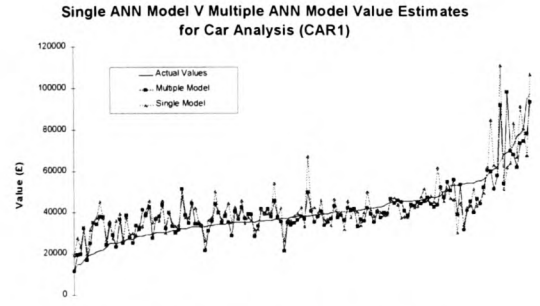
Amenities Analysis AMEN2 (Errors)

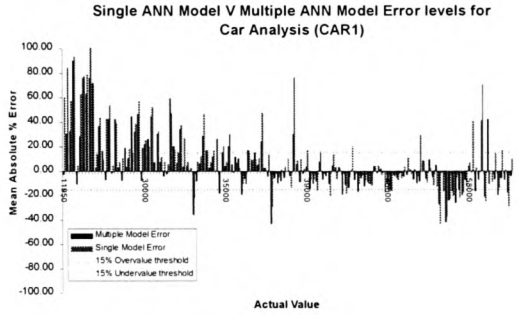




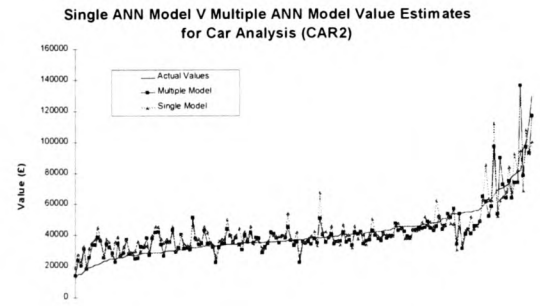
Amenities Analysis AMEN3 (Predictions) ***Amenities Analysis AMEN3 (Errors)***

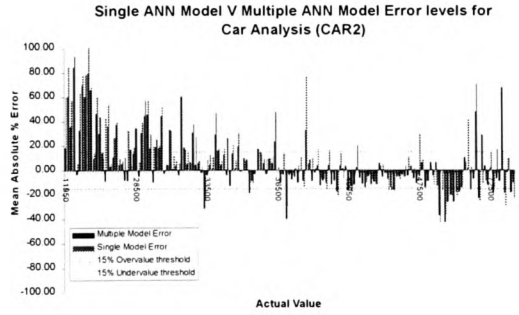
A3.5 Car Availability Analysis



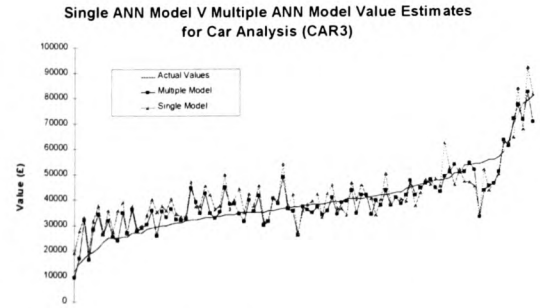


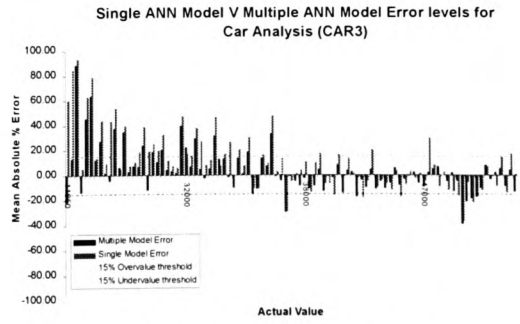
Car Avail. Analysis CAR1 (Predictions) ***Car Avail. Analysis CAR1 (Errors)***





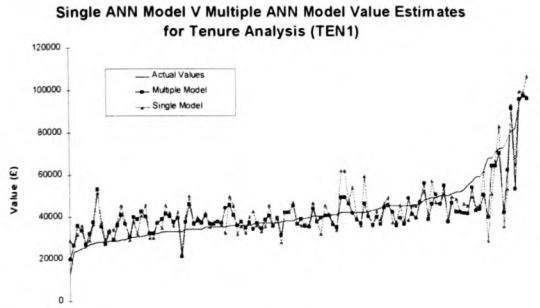
Car Avail. Analysis CAR2 (Predictions) ***Car Avail. Analysis CAR2 (Errors)***



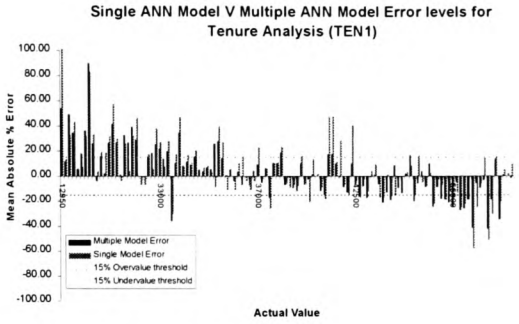


Car Avail. Analysis CAR3 (Predictions) ***Car Avail. Analysis CAR3 (Errors)***

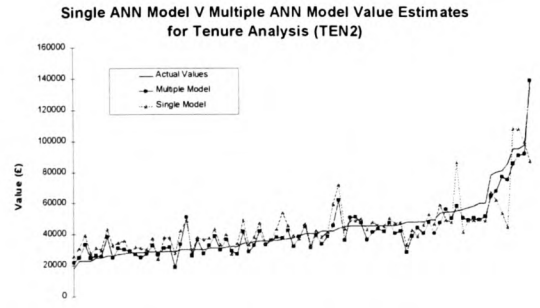
A3.6 Tenure Analysis



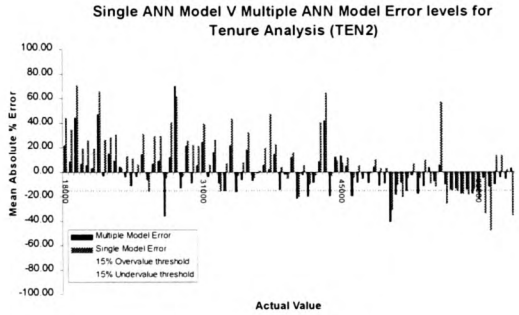
Tenure Analysis TEN1 (Predictions)



Tenure Analysis TEN1 (Errors)

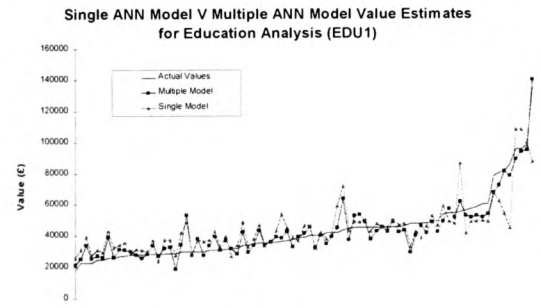


Tenure Analysis TEN2 (Predictions)

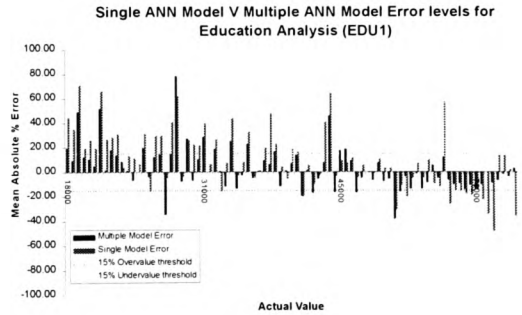


Tenure Analysis TEN2 (Errors)

A3.7 Education Analysis

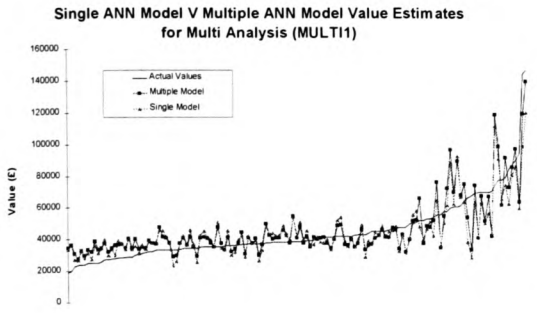


Education Analysis EDU1 (Predictions)

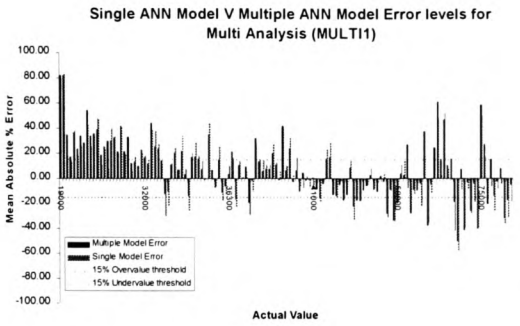


Education Analysis EDU1 (Errors)

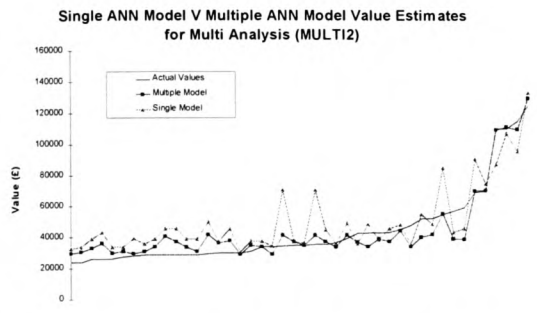
A3.8 Multi Category Analysis



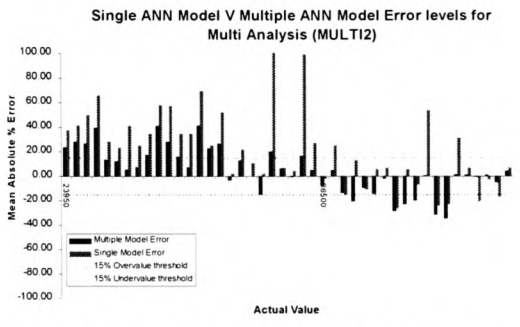
Multi-Category Analysis MULTI1 (Pred)



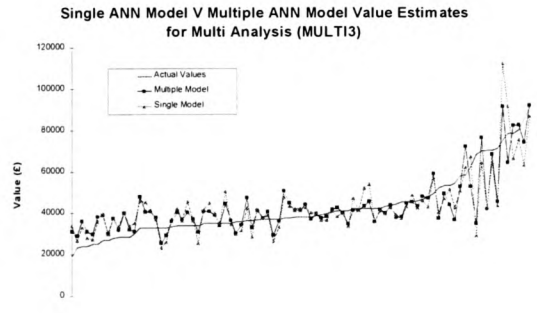
Multi-Category Analysis MULTI1 (Errors)



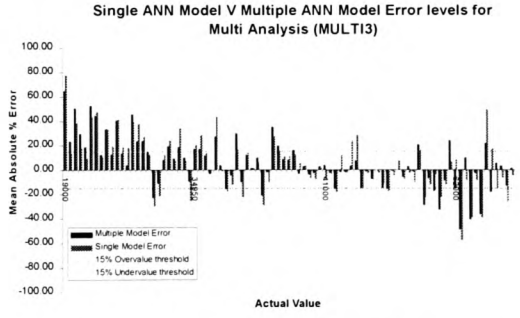
Multi-Category Analysis MULTI2 (Pred)



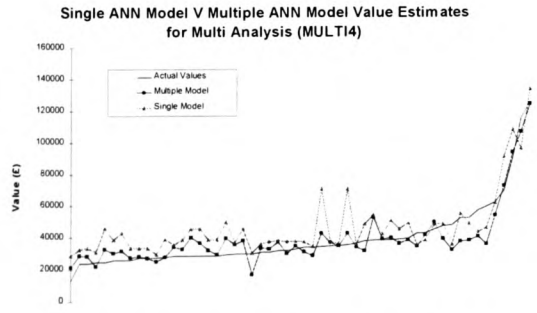
Multi-Category Analysis MULTI2 (Errors)



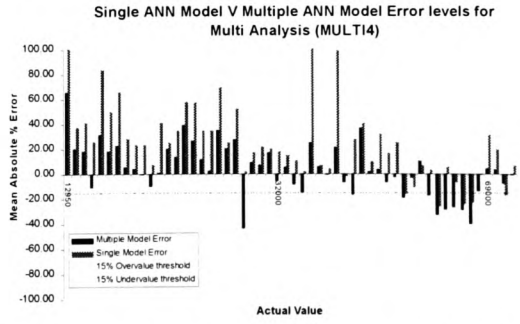
Multi-Category Analysis MULTI3 (Pred)



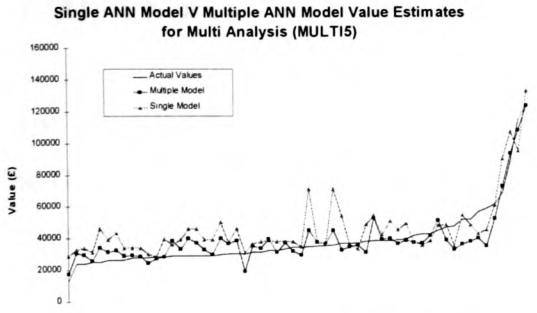
Multi-Category Analysis MULTI3 (Errors)



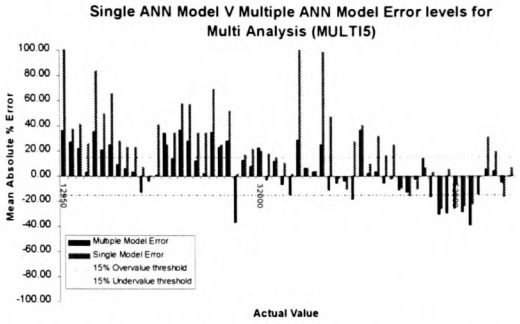
Multi-Category Analysis MULTI4 (Pred)



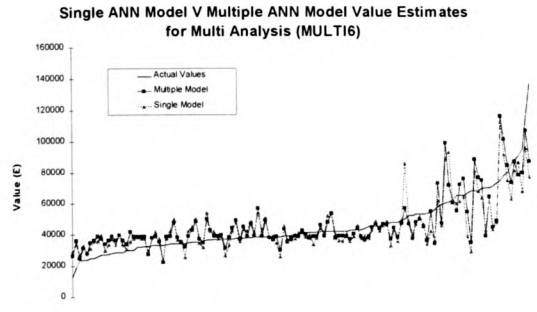
Multi-Category Analysis MULTI4 (Errors)



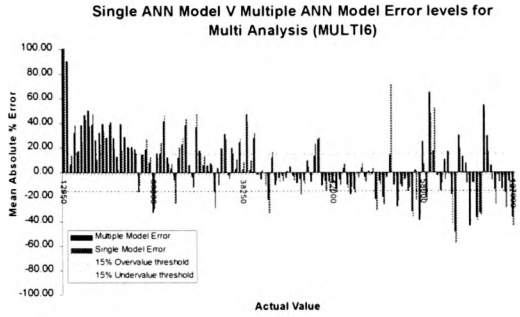
Multi-Category Analysis MULTI5 (Pred)



Multi-Category Analysis MULTI5 (Errors)



Multi-Category Analysis MULTI6 (Pred)



Multi-Category Analysis MULTI6 (Errors)

APPENDIX 4 - FURTHER INFORMATION AND RESOURCES FOR ALGORITHMS EMPLOYED

A4.1 Gamma Test

The Gamma test was distributed freely as an academic resource from the Computer Science Dept at the University of Wales, Cardiff. However, towards the end of 1998, the Gamma Test has moved from an academic resource into a commercial project with latest product information available from:

http://www.cs.cf.ac.uk/User/P.J.Durrant/winGamma.html

Useful publications:

Stefansson, A, Koncar, N, Jones, AT, 1997, A Note on the Gamma Test, Journal of Neural Computing and Applications, Vol.5 No. 3, Springer Verlag

Koncar N: Optimisation Methodologies for Direct Inverse Neurocontrol, Ph.D. Thesis, Department of Computing, 180 Queen's Gate London, SW7 2XZ, U.K, 1997.

A4.2 Rule Extraction from Kohonen Self Organising Map

The following is a simplified version of the algorithm used to extract rules from the Kohonen self-organising map.

```
For each cluster
        Start new Ruleset
        Ruleset = {Ruleset for Cluster Cluster_Name}
        For each attribute
                Calculate:      mean value              as MEAN_VALUE
                                Upper Quartile          as Q3
                                Lower Quartile          as Q1
                Rule = {IF Attribute_Name has mean = MEAN_VALUE
                        AND Attribute_Name between Q1 and Q3}
                Ruleset = Ruleset + Rule
        Next attribute
        Ruleset = Ruleset + {THEN example belongs to Cluster Cluster_Name}
Next Cluster
```

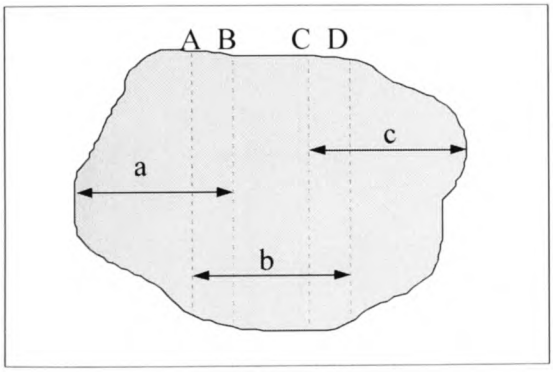
A4.3 Rule Extraction from Genetic Algorithm Chromosome Encoding

The extraction of rules from the genetic algorithm describing the composition of the homogeneous subsets involves decoding each 'elite' chromosome string representing homogeneous subsets into its original 'soft-partition' thresholds. The

following is a simplified version of the algorithm used to decode the chromosomes into 'expert system' type rules for chromosomes created using 3 soft-partitions:

```
For each Chromosome
    Start new Ruleset
    Ruleset = {Ruleset for Chromosome & Chromosome_Name}
    For each Attribute Sub-section (3 bits)
        Select Case (Sub-section)
            WHEN = [0,0,0]  THEN  Rule = Rule + {Attribute_Name = All Values)
            WHEN = [1,0,0]  THEN  Rule = Rule + {Attribute_Name < Threshold B)
            WHEN = [0,1,0]  THEN  Rule = Rule + {Attribute_Name > A AND Attribute_Name <= A)
            WHEN = [0,0,1]  THEN  Rule = Rule + {Attribute_Name > C)
            WHEN = [1,0,1]  THEN  Rule = Rule + {Attribute_Name <= A AND Attribute_Name > C)
            WHEN = [1,1,0]  THEN  Rule = Rule + {Attribute_Name <= D)
            WHEN = [0,1,1]  THEN  Rule = Rule + {Attribute_Name > A)
            WHEN = [1,1,1]  THEN  Rule = Rule + {Attribute_Name = All Values)
        Ruleset = Rulest + Rule
    Next Attribute Sub-section
    Ruleset = Ruleset + {THEN example belongs to Cluster Cluster_Name}
Next Chromosome
```

The thresholds used in the example code are based on the 2 soft-partitioning method as shown in the figure below



Example Census Variable : C_1

[a,b,c]	EDs Selected
[0,0,0]	All
[1,0,0]	C_1 <= B
[0,1,0]	A < C_1 <=D
[0,0,1]	C_1 > C
[1,0,1]	C_1 <= B or C_1 > C
[1,1,0]	C_1 <= D
[0,1,1]	C_1 > A
[1,1,1]	All

The original thresholding functions are show in (1) and (2) below.

$$\text{LowerBound}_i = (i-1)*(\beta-\alpha) + Min \quad (1)$$

$$\text{UpperBound}_i = LowerBound_i + Width \quad (2)$$

Where: i: Partition Number

$Width = 2\alpha + \beta$

$\beta = \dfrac{Max - Min}{n}$

$\alpha = Overlap\ Ratio * \beta$

Max: Maximum Value for Feature

Min: Minimum Value for Feature

APPENDIX 5 -PREDICATE LOGIC SYNTAX

Predicate Logic Symbol	Description
$\mathbb{Z}$	Real Number set
$\notin$	Non membership
$\mathbb{P}$	Power set
$\in$	Membership
$\emptyset$	Empty set
$\forall$	Universal quantifier (for all)
$\mathbb{N}$	Natural number set
$\bullet$	Then
.	Component
$\cup$	Set Union
$\cap$	Set Intersection
$\rightarrow$	One-to-one function
|	Such that
$\neq$	Inequality
=	Equality
$\vee$	Logical Or
$\wedge$	Logical And
:	Of type

Useful reference:

Spivey, J.M., 1992, The Z Notation: a Reference Manual, Prentice Hall International, Second Edition.

APPENDIX 6 - PUBLISHED PAPERS

A6.1 Lewis, O.M., Ware, J.A. and Jenkins, D.H.

"A Novel Neural Network Technique for the Valuation of Residential Properties", Journal of Neural Computing and Applications, Vol. 5, Springer Verlag, 1997.pp 224-229.

A Summary of this paper was also presented at the International Conference for Artificial Neural Networks and Genetic Algorithms, at the University of East Anglia, Norwich, 1997.

A 'work in progress' version of this paper was presented to the Institute for Quantitative Investment Research Autumn Seminar in the Bath Spa Hotel, Bath in August 1996.

A Summary of this work also formed part of an E.S.R.C. report together with other material relating to residential property valuation with the following reference:

Gronow SA, Ware JA, Jenkins DH, Lewis OM and Almond NI, 1996, A Comparative Study of Residential Valuation Techniques and the Development of a House Value Model and Estimation System. ROPA end of award report (Available as an occasional paper from University of Glamorgan).

A Novel Neural Network Technique for the Valuation of Residential Property

O.M. Lewis[1,2], J.A. Ware[1], D. Jenkins[2]

[1]School of Accounting and Mathematics; [2]Centre for Research in the Built Environment. University of Glamorgan, Trefforest, Mid Glamorgan, UK.

A number of published studies have investigated the application of neural network technology to residential property appraisal. The majority of these studies have concentrated on homogeneous areas (that is areas where properties are subject to the same environmental and locational factors). This is generally done to restrict the data set to one local sub-market. However, the models created are specialised and not locationally portable. This paper presents a methodology, which builds on research reported by James [1], in which a Kohonen map is used to uncover sub-markets within a large data set that are subsequently independently used to train a series of back-propagation networks. (The paper also introduces a novel boundary detection algorithm for a Kohonen self organising map.) The study concludes that by modelling possible sub-markets an acceptable accuracy over a heterogeneous area can be achieved. The work presented in this paper is funded via a Realising Our Potential Award under the auspices of the ESRC.

Keywords: Kohonen Feature Map; Class Discriminants; Gamma Test; Residential Property Appraisal; Back Propagation; Variance Estimation

Introduction

In the UK, and indeed in many countries, the valuation of residential property is based on the method of direct capital comparison. However, the principal weakness of this method of valuation is the problem of obtaining suitable comparable properties as evidence [2]. Given this weakness, research has been carried out on directly calculating the value of a property from its locational and physical attributes [2,3,4,5]. Many of these research studies have considered the application of neural networks to residential property appraisal [4,5], with the majority of studies using data from a homogeneous area (i.e. an area where all properties are subject to the same environmental and locational forces). This approach is taken as the valuation function can become very elaborate when spread across a heterogeneous area.

This however leads to neural network models which are not locationally portable. Nevertheless, the studies have reported a high level of success, with average absolute percentage error levels of between 5 and 7.5% not being uncommon [4, 6].

Adair [7] hypothesises that sub-markets can be identified by stratifying the market into increasingly homogeneous subsets. Using the hypothesis that a heterogeneous area consists of many homogeneous areas, the authors postulate that a heterogeneous area can be modelled indirectly using many back propagation networks, trained on subsets of the parent data set. To use such a system to predict the value of a previously unseen property, requires a method of selecting the appropriate neural network model. This paper describes such a method, in which a Kohonen feature map is used both to identify groupings within the parent data set and to act as a panel judge to decide which neural network model to select when asked to give a valuation.

Overview of Methodology

The methodology involves training a Kohonen network on historical data, collected from approved mortgage transactions over a number of years. The data set from each significant grouping - formed by the Kohonen network - become the training set for a back-propagation network. Each back-propagation network is trained using a save-best approach. This involves periodical testing of the network with a validation set and, if the current state produces better results than the previous best, the current state becomes the new best state. At the end of training the current best state is adopted as the network to be used during day-to-day operations. Figure1 provides an overview of the methodology.

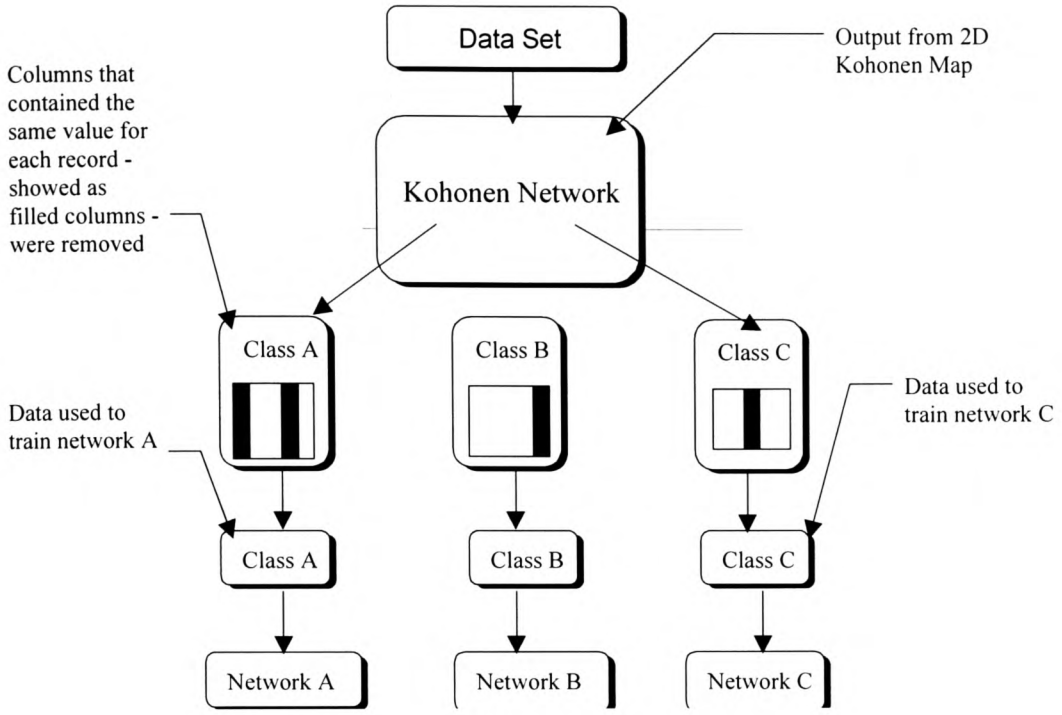


Figure 1 - An Example Feature Map for a Trained Kohonen Network.

The advantage of using the Kohonen self organising map for this application is that it can identify clusters within the parent data set that are difficult to achieve using simple sort procedures. However, it is sometimes difficult to identify class boundaries within a trained Kohonen map [1], and this in turn leads to problems in generating training sets for the back propagation networks. For example, consider the feature map shown in Figure 2; there appear to be five classes within the data set, but there are regions of uncertainty relating to the boundaries of each cluster. The boundaries could be estimated visually, but no doubt at the expense of accuracy.

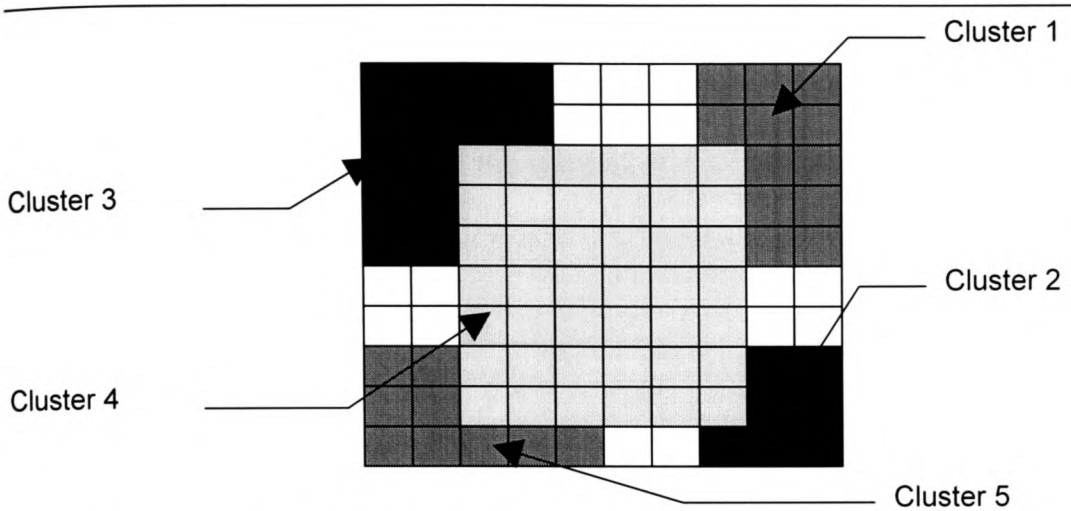


Figure 2 - An Example Feature Map for a Trained Kohonen Network.

To overcome this problem a simple method of identifying class boundaries or discriminants can be used, which relies on the fact that the Kohonen network clusters primarily on binary features. For example, if the type of house is represented using binary inputs, the Kohonen network will tend to cluster records according to this variable. Boundaries between adjacent clusters on a 2D map can then be found by inspecting the records mapped to each node and grouping together nodes that contain the same classification values. However, this level of clustering can be achieved using a multi-level sort procedure. In essence, the binary representation of the data will directly influence the make-up of the resulting clusters and possibly the homogeneity of the data sets.

If the data are represented using continuous inputs, the clusters formed by the Kohonen map would provide more generalised classes which would be difficult to achieve using a sort procedure. However, the inspection method would no longer identify class boundaries as the similarities between records would not always be apparent. Clearly, before meaningful training data sets can be formed, the problem of discerning effective class boundaries in a Kohonen feature map must be addressed. Ideally, the network adaption rule should cluster similar inputs and clearly distance individual clusters. Zurada [8] explains "One possible network adaption rule is: A pattern added to the cluster has to be closer to the centre of the cluster than to the centre of any other cluster". Using this rule, each node can be examined

and the distance from the surrounding centroids[5] can be calculated. The subject node can then be added to the nearest cluster. Figure 3 illustrates a hypothetical situation where it is unclear where to draw the boundaries around clusters on a Kohonen map. (The numbers shown in each node box represent the number of input vectors mapped to each individual node.)

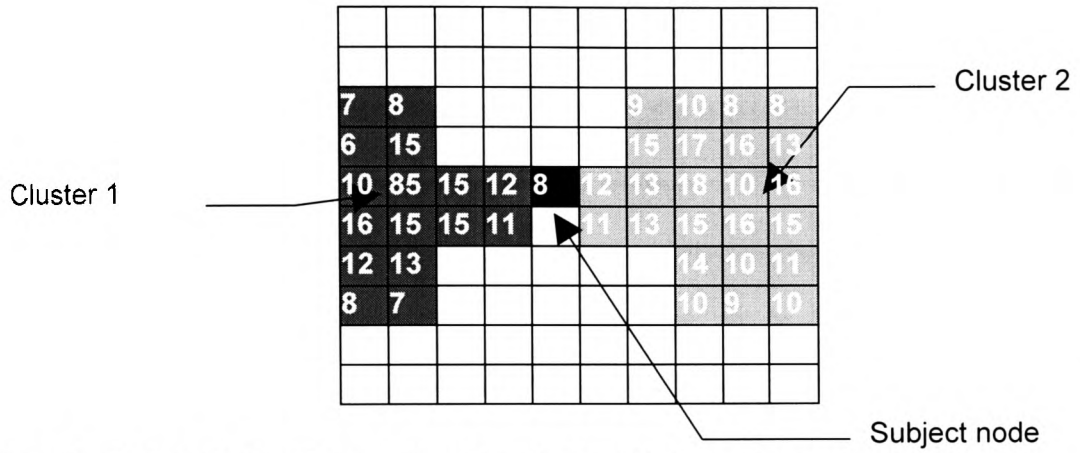


Figure 3 - A Hypothetical Feature Map for a Trained Kohonen Network.

By simply calculating the Euclidean distance of the subject node from the two centroids, the subject node can be assigned to the cluster which is closest [1] - for example Cluster 1. However, in this application, which aims to generate useful training data sets, the formation of a class boundary for Cluster 1 (including the subject node) may dramatically increase the variance of the training data. This increase will in turn reduce the potential accuracy of the back propagation model. In the example, it may have been better to exclude the subject node from either of the clusters, and deem the vectors mapped to the subject node as either being outliers or a separate cluster.

[5] A centroid is taken to be a node, outside any known cluster boundaries, that has the largest number of input vectors mapped to it.

In addition to identifying boundaries around input clusters, it is also important in this application for input clusters to be matched by corresponding output clusters. In terms of residential property appraisal, if, for example, the Kohonen map has clustered residential properties from two different locational areas, it is reasonable to expect similar types of houses from each area to have a similar property value. Figure 4 gives examples of useful and useless clustering.

In summary, for this methodology to be successful, the following is required:

- class boundaries must be identified around clusters formed by the Kohonen feature map over the input space that exclude outliers and nodes from neighbouring clusters, and;
- only 'good' clusters (see Figure 4a) should go on to form training data sets for subsequent back propagation models.

Faced with this problem, the authors decided to investigate a recently published variance estimation routine known as the Gamma test [9] with the expectation that this would address these two requirements.

The Gamma test attempts to estimate the best mean square error that can be achieved by any smooth modelling technique using the data. If y is the output of a function then the Gamma test estimates the variance of the part of y that cannot be accounted for by a smooth (differentiable) functional transformation. Thus if $y = f(x) + r$, where the function f is unknown and r is statistical noise, the Gamma test estimates Var(r).

Var(r) provides a lower bound for the mean squared error of the output y, beyond which additional training is of no significant use. Therefore, knowing Var(r) for a data set allows prediction beforehand of what the MSE of the best possible neural network trained on that data would be. [9]

The idea that the performance of a neural network can be predicted for a particular data set before engaging in any modelling processes seemed to fit in well with the problem at hand. After some initial research to investigate optimum parameters to supply to the Gamma test, and the significance to place upon each element of the output from the test, the authors began to investigate how the Gamma test could be employed to identify useful training clusters within a Kohonen map.

The algorithm produced, which is shown in Figure 2, attempts to identify useful clusters by selecting a centroid and adding neighbouring nodes - where the addition of a node increases the variance significantly it is subsequently removed. This process iterates until the cluster size is maximised within a specified variance threshold.

```
Select the node with the largest record count.
Using the Gamma test estimate the variance for the data in the selected node
Set Number_of_Neighbours = 8
For i = 1 to Number_of_Neighbours
-add data from node i to the cluster of previously examined data
-run gamma test on new cluster
-Decide (based on a variance threshold ) whether to include or exclude
 data  from node i in cluster
Number_of_Neighbours = Number_of_Neighbours + 8*
Repeat 4 until inclusion of all additional neighbours passes threshold.
Record the boundaries of this cluster.
Select a node outside the recorded boundaries with the largest record count.
Repeat 2 to 7 until all nodes with large record counts have been investigated.
```

Figure 2 - Boundary Detection Algorithm.

As the Gamma test estimates the variance within a data set, prior to boundary detection the variance threshold can be set, and therefore the level of performance of a neural network trained on the sub-clusters can be predicted.

Testing the Methodology.

A database containing information on residential property transactions during the period January 1993 to December 1995 was selected to test the methodology. There are 51

attributes in the original database. However, a number of these have either constant values or free form text fields which are difficult to recode and were therefore removed. A description of the database used for this study is shown in Table 1.

Table 1 - A description of the database.

Name of Field	Example Value
Street Name	Newport Road
District or Village	Roath
Unit	1 - 6
Unit Type	Mid terraced etc.
Unit Size	Area M^2
Valuation Date	950512
Main Heating	Full, Partial, None
Number of Bedrooms	1 - 8
Age in Years	0 - 500
Number of Garages	0 - 2
Value	10,000-255,000

A 10 by 10 Kohonen map was used to find the groupings in a historical data set containing 990 records. Value, the output feature, was omitted from the Kohonen training set. Using a combination of the boundary detection methods previously described, the data were found to contain eight groups. The records from each group were examined and common features within the group removed. Obviously, as the data set is partitioned into classes, the classes contain only a portion of the original 990 records. However, this is accompanied by a decrease in the number of dimensions - as constant columns were removed.

The Impact of the Methodology on Prediction Accuracy

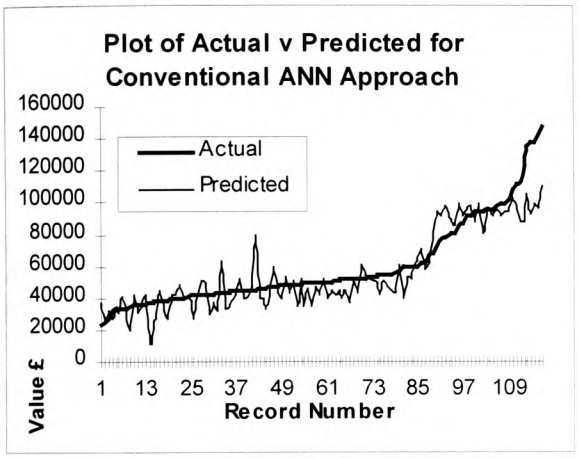
In order to provide a bench mark for analysing the methodology, a single back-propagation neural network was trained on the whole data set. After training the ability of the network to appraise residential properties with known values was tested. The results of this test are shown in Figure 2(a); the graph shows the actual and predicted values for 117 test properties (for ease of interpretation the properties have been ordered according to actual value). Table 2 illustrates the results achieved using the described methodology on the same 117 properties in the test set.

* Additional constraints required to manage Kohonen map perimeters.

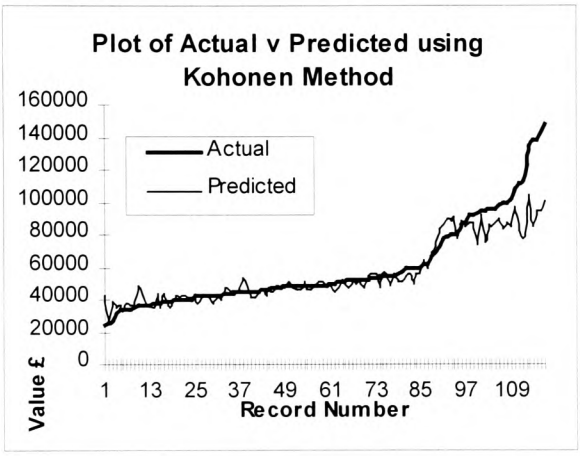
Table 2 - Results achieved for the test set.

	Conventional ANN Method	Kohonen Method
Mean absolute % error	18%	8%
% of Records with an error > 10%	74%	22%
Minimum absolute % error	0%	0%
Maximum absolute % error	310%	49%

Figure 2b shows the improvement in accuracy using the new method over the conventional approach (the data used for Figure 2a and Figure 2b have been sorted in ascending actual property value).



(a) A graph of actual and predicted value gained using a conventional neural network approach.



(b) A graph of actual and predicted value gained using the Kohonen/BP hybrid methodology.

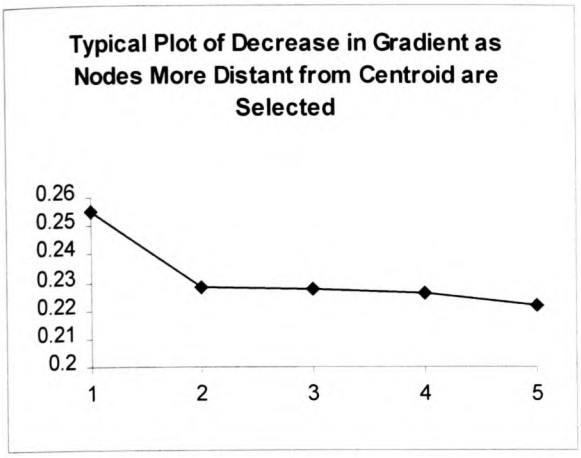
Figure 2

The Performance of the Boundary Detection Algorithm

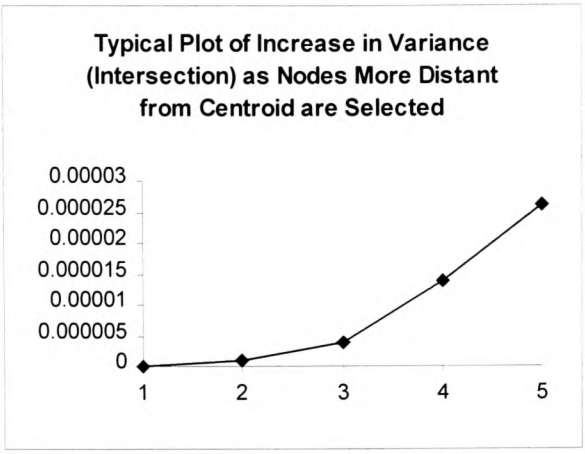
Although this research did not initially set about to address the problem of detecting boundaries in a Kohonen feature map, it is worth summarising the main observations made:

- to take full account of the variance within a data set, the number of nearest neighbours to select for the Gamma test should be carefully chosen;
- a gradient of 1 and an intersection of 0 describes a data set with negligible variance;
- both the gradient and the intersection can give an indication of variance.

Figure 7 clearly shows how both the intersection (Figure 3a) and the Gradient (Figure 3b) change as the variance increases by adding nodes which are progressively further away from the centroid.



(a)



(b)

Figure 3

(a) A Plot of Least Squares Intersection Increasing, Absolutely from Zero, as Variance Increases.

(b) A Plot of Least Squares Gradient Decreasing as Variance Increases.

Whilst the authors are cautiously encouraged by the preliminary results obtained using this algorithm, further work is being undertaken to examine both the boundary detection algorithm and the Gamma test on which it is based.

Conclusion

It is evident, from the results obtained, that the methodology proposed in this paper compares very favourably with the more conventional neural network approach. An average increase in prediction accuracy of 10% was achieved using the new method over the conventional approach. This implies that the original data set either contained more than one underlying function (pattern) [1] or the function was too elaborate to be modelled using a single back propagation network. Moreover the Kohonen network can discern different classes within the data, which when independently modelled yield a greater predictive accuracy than those computed for the original data set [1]. In addition to this, analysis suggests the Kohonen step can be applied to subsets of the data (for example Class A) to create even more accurate sub-models.

From the work outlined in this paper, the authors have concluded that the techniques described may usefully be applied to other data sets. The authors are currently in the process of analysing data from the 1991 Census, using the methods described, with the aim of grouping together enumeration districts that contain the same valuation functions. This will allow the modelling process to move from specialised models to locationally portable systems.

REFERENCES

James, H. 1994. An 'Automatic Pilot' for Surveyors. RICS Cutting Edge Conference

Worzola, E, Lenk, M, Silva, A. 1995. An Exploration of Neural Networks and Its Application to Real Estate Valuation, Journal of Real Estate Research: 185,201.

Adair, A S, and McGreal, S 1987, The Application of Multiple Regression Analysis in Property Valuation, Journal of Valuation, Vol. 6, 57-67.

Evans, A., James, H., Collins, A. 1992. Artificial Neural Networks: an Application to Residential Valuation in the UK, Journal of Property Valuation & Investment : 11,195-204.

Do, Q. & Grudnitski, G. 1992. A Neural Network Approach to Residential Property Appraisal, The Real Estate Appraiser, 38-45.

Borst, R.A. 1991. Artificial Neural Networks: The Next Modelling/Calibration Technology for the Assessment Community ?, Journal of Property Tax : 10,1,69-94.

Adair, A S, Berry, J N, McGreal, W S, 1996, Hedonic Modelling, Housing Submarkets and Residential Valuation, Journal of Property Research, Vol. 13, 67-83.

Zurada, J M, 1992, Introduction to Artificial Neural Systems, West Publishing Company (ISBN 0-314-93391-3) p58.

Koncar N: Optimisation Methodologies for Direct Inverse Neurocontrol, Ph.D. Thesis, Department of Computing, 180 Queen's Gate London, SW7 2XZ, U.K, 1997.

A6.2 Lewis, O.M., Ware, J.A. and Jenkins, D.H.

"A Novel Neural Network Technique for Modelling Data Containing Multiple Functions", in Computational Intelligence - Theory and Applications, ed. Bernd Reusch, (Lecture Notes for Computer Science Series Vol. 1226), Springer Verlag, ISBN 3-540-62868-1, pp 141-149.

A Novel Neural Network Technique for Modelling Data Containing Multiple Functions

Owen M. Lewis & J.Andrew Ware

Division of Mathematics and Computing, University of Glamorgan
Pontypridd, Mid Glamorgan, UK.

Abstract.

Increasingly neural network techniques are being applied to a wide range of pattern recognition and classification problems. However, there is often insufficient information available to facilitate optimal operation. This problem can lead to a situation where the data exhibits signs of containing multiple underlying functions. For example, if location is not included as a feature when modelling residential property appraisal, the data will appear to map across more than one underlying function. The methodology proposed in this paper uses a form of data stratification to overcome this problem. The premise followed is that it is better to produce multiple models that are specific to - and accurate within - certain scenarios, rather than a single model that is too general and therefore inaccurate.

Introduction

One of the major unresolved issues affecting the use of neural networks is the selection of a suitable input vector that will facilitate the correct determination of the output vector. Moreover, users of neural networks often have available an insufficient subset of the information that would enable effective modelling of the mapping function between input and output space.

When there is sufficient *a priori* knowledge, selecting input features is a relatively simple task. For example, when attempting to model house prices, common sense dictates that there is no need to include 'door colour' in the input vector. Similarly, it is intuitive that if the 'number of bedrooms' or 'location' is missing from the input vector then the model will only have limited use in prediction.

When input space does not include the full complement of features that would enable effective mapping to take place, then it resembles a situation in which the input space contains more than one underlying function. For example, two identical houses may differ in value by £50,000 because of their location - if locational information is not available then a single function cannot model the scenario successfully - whereas two separate functions might. Therefore, this paper considers an input space containing missing features to be equivalent to an input space generated by multiple functions.

The problem of modelling data containing multiple underlying functions can be ameliorated by stratifying the data set into a number of subsets that each contain a single underlying function. Consider the simple following example: a data set containing records produced by the following functions:

$$x = 6a + 5b - c$$
$$y = 4a - b + 3c$$
$$z = 2a + 3b + 7c$$

Table 1 shows an example of such a data set.

Table 1 - An extract from the generated data set.

a	b	C	result
1	2	3	13
1	2	3	11
1	2	3	29
:	:	:	:
:	:	:	:
:	:	:	:
2	2	2	20

If a complete data set of this type was analysed using either multiple regression analysis or neural networks, the ability to predict the output given the values of a, b and c would be poor. However, if the records could be split up into three separate data sets (a data set for those records produced by equ.1, a data set for those records produced by equ.2 and finally a data set for those records produced by equ.3) then the problem could be solved.

In order to understand how to achieve this data stratification a number of scenarios related to clustering in input and output spaces are now considered. Figure 1 shows two of these scenarios.

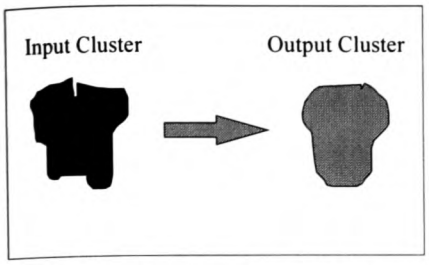


Figure 1a - Useful Input Clusters

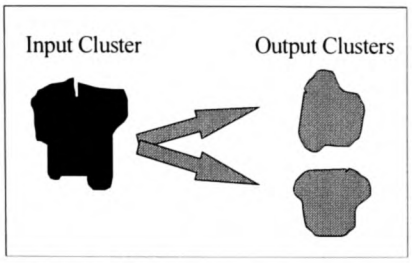
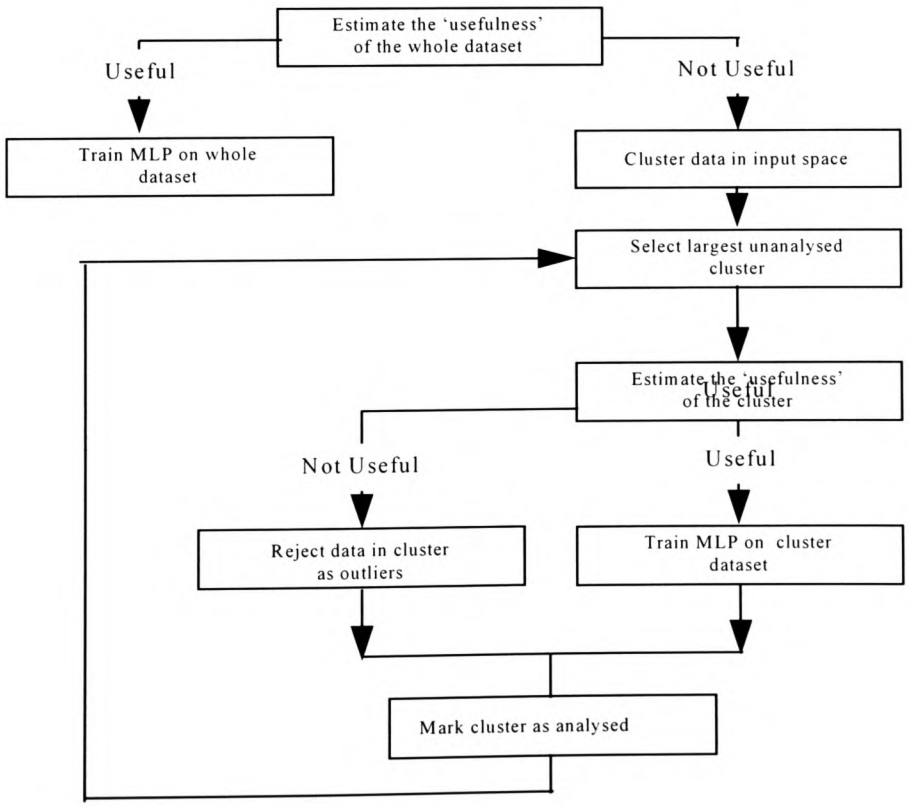


Figure 1b - Useless Input Clusters

Figure 1a shows a cluster of similar input vectors. When the corresponding data in output space is examined all the examples describe similar output values. For example, if the input cluster describes houses that have three bedrooms; semi-detached; under 2 years old - then the cluster in output space says they all have similar property values. Conversely, Figure1b shows a situation where the data can only be modelled using two functions.

In order to determine whether the whole of the input and output space is suitable for being modelled with a single MLP - or whether data stratification needs to be performed - the following algorithm can be applied:



(1.0)

The problem now is to estimate the 'usefulness' of a given cluster. This can be achieved using a method based upon a variance estimation routine known as the Gamma test (Koncar 1997).

The Gamma Test

The Gamma test attempts to estimate the best mean square error that can be achieved by any smooth modelling technique using the data. If y is the output of a function then the Gamma test estimates the variance of the part of y that cannot be accounted for by a smooth (differentiable) functional transformation. Thus if $y = f(x) + r$, where the function f is unknown and r is statistical noise, the Gamma test estimates Var(r).

Var(r) provides a lower bound for the mean squared error of the output y, beyond which additional training is of no significant use. Therefore, knowing Var(r) for a data set allows prediction beforehand of what the MSE of the best possible neural network trained on that data would be.

Applying this test to the data set shown in table 1 would obviously result in a high variance. However, the Gamma test provides a method of determining the quality of the data stratification - a good stratification technique will result in a low value of Var(r) for each subset.

Methodology

In order to implement algorithm (1.0) a Kohonen Self Organising map can be used to cluster the data in the input space. The framework shown in Figure 2 illustrates the methodology.

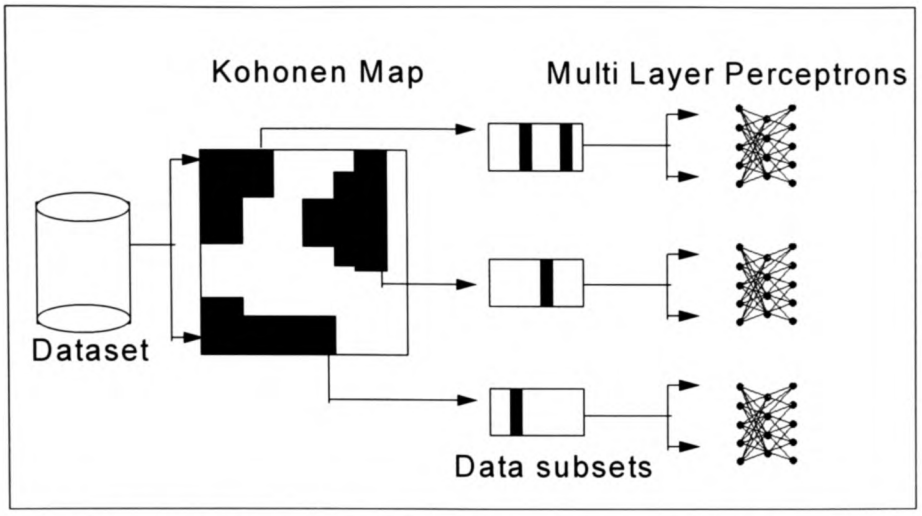


Figure 2 - Methodology used for stratification of data. **Once the Kohonen network has converged, the data from each significant grouping becomes the training set for a MLP. Each MLP is trained using a save best approach. This involves periodical testing of the network with a validation set and, if the current state produces better results than the previous best, the current state becomes the new best state. At the end of training the current best state is adopted as the network to be used during day-to-day operations.**

The methodology divides into three abstraction levels:

- Cluster level
- Node Level
- Record Level

Data stratification is achieved at cluster level or at node level, depending on the ease at which cluster boundaries can be determined. The record level gives an indication of outliers.

Cluster Level Analysis

This can be achieved thus:

```
Identify Cluster boundaries in Kohonen map
For every cluster
        Place records mapped to  cluster   into a file
        Apply Gamma test to data in the    file
        If Var(r) <= some Threshold  then
           Use data file as the training set for a MLP
        else Process at Node Level
```

(2.0)

This level of abstraction is the least computationally intensive as it only requires one pass of the Gamma test for each cluster. The disadvantage with this method is that it is often difficult to identify boundaries between adjacent clusters on a Kohonen self organising map. Consider the feature map shown in Figure 3.

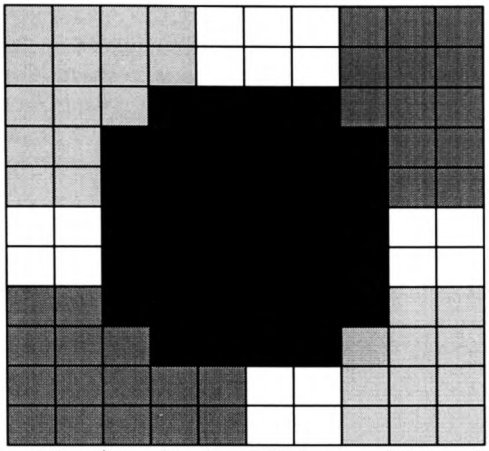


Figure 3 - An Example Feature Map for a Trained Kohonen Network.

There appears to be five classes within the data set, but there are regions of uncertainty relating to the boundaries of each cluster. The boundaries may be estimated visually but this may result in a loss of accuracy.

If binary input variables are used, class boundaries can be found by inspecting the records mapped to each node and grouping together nodes that contain the same values. Unfortunately, using binary inputs with a Kohonen network, often gives groupings that are no better than those achieved using a simple sort procedure. However, if continuous input variables are used to represent the data, this inspection method will no longer work as the similarities between records will not always be apparent. A solution to this problem is to examine the Euclidean distance from the subject node to the nearest identified centroids, and attribute the subject node to the cluster that it is closest to (Zurada 1992).

Node Level Analysis

At this abstraction level, the methodology attempts to identify useful clusters by selecting a centroid and adding neighbouring nodes - where the addition of a node increases the variance significantly it is subsequently removed. This process iterates until the cluster size is maximised within a specified variance threshold. This is achieved thus:

```
Number_of_clusters:=0
While there are nodes to cluster

        number_of_clusters : = number_of_clusters + 1
        Select the unclustered node with the largest record count
        Apply Gamma test to estimate the variance for the data in the selected node

        If Var(r) <= Threshold then Nodes_of_interest:=None
       (Cluster includes only the data from selected node)

            For each unclustered node immediately surrounding selected node
                Add data from unclustered node to the cluster
                Run gamma test on cluster
                If Var(r) <= Threshold then
                        Add unclustered node number to nodes_of_interest
                Else
                        Remove data from the unclustered node from the cluster
                While nodes_of_interest <> None
                        Select c_node from nodes_of_interest
                        Remove c_node from nodes_of_interest
                        For each unclusterd node immediately surrounding c_node
                                Add data from unclustered node to the cluster
                        run gamma test on cluster
                                If Var(r) <= Threshold then
                                        Add unclusterd node  to nodes_of_interest
                                Else
                                        Remove data from the node from cluster

                                Record the boundaries of this cluster

        else Process at Record Level as node has too high variance to train an ANN
```

(3.0)

This algorithm identifies useful clusters on a 2D Kohonen map. The boundary detection algorithm for a 1D Kohonen map is very similar except neighbouring nodes are selected progressively further away from the left and the right of the centroid node.

This level of analysis is more computationally intensive than algorithm (2.0) as it require $m*\sum n_i$ passes of the Gamma test, where 'i' is the number of nodes investigated for cluster 'n' for a Kohonen map containing 'm' clusters.

If using either the cluster level analysis (2.0) or the node level analysis (3.0), useful clusters have been identified, it is then possible to train an independent MLP on each subset. The Kohonen map is then used to select the appropriate MLP on which to predict the value of a previously unseen example. The resulting system is closely related to a panel judgement system.

However, if the methods described in (2.0) and (3.0) have still resulted in poor training sets (useless clusters) then the analysis is taken to the most detailed abstraction level, that is the record level.

9.1 Record Level Analysis

The record level analysis is the most computationally intensive. The purpose of this level of the methodology is to identify data subsets from examples that have mapped to the same node on the Kohonen map. This facilitates extraction of outliers from a data set as well as giving some indication as to the examples that require additional features.

The algorithm developed for this level of analysis is very similar to that shown in (3.0). However, this time it is sets of records that are iteratively analysed using the Gamma test. This is achieved thus:

```
For each node in the Kohonen SOM
        Apply Gamma test to estimate the variance for the data in node
        If Var(r) > Threshold then
                For each record at node
                        Remove record from data set
                        Apply Gamma test to estimate the variance for the data in node
                        If New Var(r) < Previous Var(r) then
                                Mark record as outlier
                else
                                Add record back into data set
        Else Proceed at Node Level
```

(4.0)

This level of analysis will identify the need for additional features and highlight records that may be classed as outliers.

Conclusions

It is envisaged that the methodology outlined will be useful in many application areas. For example, it has been tested using property data from the UK with the objective of predicting a property price given a set of attributes. The methodology compared very favourably with neural network models trained on the whole data set. The results show an average increase in prediction accuracy of 10% (Lewis et al. 1997). Similar results were achieved when using the methodology as a predictor in a digital elevation model data compressor (Ware et al. 1997).

References

Koncar N: Optimisation Methodologies for Direct Inverse Neurocontrol, Ph.D. Thesis, Department of Computing, 180 Queen's Gate London, SW7 2XZ, U.K, 1997.

Lewis OM, Ware JA, Jenkins DH: A Novel Neural Network Technique for the Valuation of Residential Property, Journal of Neural Computing and Applications, Springer Verlag, 1997.

Zurada, JM: Introduction to Artificial Neural Systems, West Publishing Company (ISBN 0-314-93391-3) p58, 1992.

Ware JA, Lewis OM, Kidner DB: A Neural Network Approach to the Compression of Digital Elevation Models, 5th GISRUK Research Conference - Leeds, 1997.

A6.3 Lewis, OM, Ware, JA and Jenkins DH 1997,

"The Use of Census Data in The Appraisal of Residential Properties Within the United Kingdom: a Neural Network Approach", 5th European Conference and Exhibition on Geographical Information Systems, Vienna

THE USE OF CENSUS DATA IN THE APPRAISAL OF RESIDENTIAL PROPERTIES WITHIN THE UNITED KINGDOM: A NEURAL NETWORK APPROACH.

C J Lowe and D Kaze
University of Glamorgan - Division of Mathematics and Computing
Pontypridd Mid Glamorgan

Summary

In recent years, a small number of research groups have investigated the application of neural network technology to residential property appraisal. The majority of these studies have concentrated on homogeneous areas (that is areas where properties are subject to the same environmental and locational factors). This is generally done to restrict the data set to one local sub-market. However, the models created are specialised and not locationally portable. This paper presents results indicating that features extracted from Census data can provide location surrogates that significantly improve prediction accuracy. The work presented in this paper is funded via a Realising Our Potential Award under the auspices of the ESRC.

Introduction

In the late 1980's, house prices in the UK soared to record heights. However, this rapid house price inflation did not reflect real property values. Eventually, in 1989, the market collapsed under its own bloated weight plunging thousands of home owners into the negative equity trap. The valuation methods used by mortgage lenders had failed to adequately assess the risk involved in what to most people is the largest financial investment of their lives.

The conventional method for valuing a residential property is the method of Direct Capital Comparison (DCC). This involves selecting properties comparable to the subject property that have been sold in an 'open market'. The valuer makes an *"allowance in money terms*[1] for any differences between the subject property and the comparable properties.

In practice, the DCC method relies on valuers' opinion and as such is *"not fact and may be consciously or unconsciously biased"*[2]. Furthermore, appraisers have *"little objective evidence"*[2] resulting in the property value being highly correlated with transaction price[3]. The DCC method is therefore weakened by the difficulty in obtaining suitable comparable properties[4] as evidence.

Many researchers are attempting to provide valuers with valuation instruments that directly calculate the value of a property from its locational and physical attributes [4,5,6]. Techniques such as multiple regression analysis are being used to develop these tools, however, it seems *"the facts underlying real estate are much too complex for the simple additive theory on which (MRA) is based"*[7]. Another technique, which overcomes this modelling restriction, is neural networks. A technique developed by artificial intelligence researchers which is rapidly growing in popularity. However, this technique is handicapped (as is MRA) by the lack of understanding of the effect location has on property value. Hence, published studies tend to model small areas where location can be removed as a constant.

The aim of the research presented in this paper was to determine whether neural networks exhibit potential to appraise residential property for large areas of the UK using locational attributes extracted from the 1991 UK Census.

Census Data

The 1991 Census provides researchers and government with the *"most authoritative social accounting of people and housing in Britain"*[8]. Comparable statistics are generated for very fine geographical areas, the smallest of which being an enumeration district (ED) in England and Wales, and an output area (OA) in Scotland. The twenty or so questions in the Census can be cross-classified to provide *"powerful statistical insights into the social conditions of the population and its housing"*[8]. Thanks to the inclusion of postcode in the 1991 Census questionnaire, it is possible to link social and economic data with housing stock data via a postcode to enumeration district table.

Ideally, to achieve a representative quality from the Census data, information should be extracted at the ED level. However, this is a labour intensive process, and should be pre-empted with a study to determine which Census features are important from amongst the 20,000 available. Current methods of determining the features that impact on a dependent variable - in both neural networks and MRA rely heavily upon a priori knowledge. Mathematical techniques exist, examples including principle component analysis and stepwise regression, however, these tend to be linear in nature and may result in a loss of information.

Substantial work in the field of geodemographics can help in this selection process. A customer targeting tool called ACORN™, developed from Census analysis, attempts to provide a *"powerful tool"*[9] to *"address the complexity of consumer markets"*[9]. Indeed, the Nationwide Building Society's house price index uses the ACORN™ classification system. However, ACORN™ is expensive for inclusion in academic studies and although similar classification systems have been developed which are more readily available[10], it is the raw data itself which undoubtedly contains the most information.

Neural Networks

Tazelaar[11] describes neural networks as *"humanity's attempt to mimic the way the brain does things in order to harness its versatility and its ability to infer and intuit from incomplete or confusing information"*. Neural networks are able to generalise from examples and have the ability to interpolate from previous learning[12]. Neural networks are often found working as pattern classifiers in areas where problem solutions are complex and difficult to specify, but which have an abundance of data from which a response can be learnt.

The most common neural network used is the multi layer perceptron trained using a back propagation algorithm. Figure 1 shows an example of such a network.

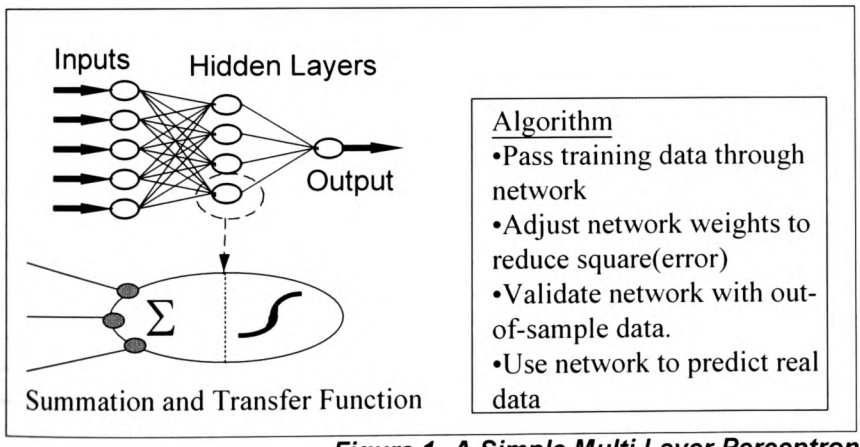


Figure 1- A Simple Multi Layer Perceptron.

For a neural network to learn the underlying function in the data, it is presented with a set of training cases known as the 'training set'. Each training case comprises an example of the inputs (features representing the case) and a target output (representing the known solution). During training the neural network attempts (by adjustment of weights) to match its predicted output with the known output.

The hidden layers allow implicit patterns, that may exist within the data, to be modelled. Each neuron employs a transfer function that governs the output value given a set of input values.

Case Study

To investigate the usefulness of Census data, the Authors extracted Census variables at the district level. The variables used were those highlighted in published literature[10,13] together with those selected using a common-sense approach - a description of the selected variables appears in the Appendix.

Census data extracted at the district level attempts to characterise a sample containing on average 2816 people in 1528 households. This allows investigation into the effect high level data imputation has on the domain modelling process.

Results

Table 1 shows the results for a number of district level trials. The tests show an improvement in accuracy at this very high level of abstraction. The best performance (a percentage accuracy increase of 2%) was achieved when all the selected Census variables were added to the data set containing details of properties from an area of South Wales.

Table 1 - Results obtained when Census data was added.

Description	Mean Abs % Error
Control case (Cardiff Data set).	28.33
Including all Census Information (As shown in Appendix)	26.36
Adding only Employment statistics to the Cardiff Data set.	26.89
Adding only People to Car Statistics to the Cardiff Data set	27.31
Adding only Occupation statistics to the Cardiff Data set	26.91
Adding only Ethnic groupings statistics to the Cardiff Data set.	27.94
Adding only Tenure statistics to the Cardiff Data set.	26.73
Adding only Property Type statistics to the Cardiff Data set.	27.24
Adding only Amenities statistics to the Cardiff Data set.	27.02

Table 2 shows the results of similar exercise performed using data extracted at the ED level using the same property and Census variables. Here the mean absolute percentage difference between the predicted value using Census data and the value returned by the valuer was 7% closer than that achieved in the absence of Census data.

Table 2- Results obtained for ED level analysis.

Data Sample	Mean Abs % Error
Sample of Cardiff Data set	20 %
Cardiff Data set & ED level Census stats.	13 %

It is the intention of the Authors to investigate methods of variable selection, using mathematical processes and also to continue studying relevant literature.

Conclusions and Future Work

It is evident, from the results obtained, that considerable improvement in accuracy is gained when Census data is added to property data for residential property appraisal modelling. An average increase in prediction accuracy of 7% was achieved when Census data was added. From the work outlined in this paper, the authors are confident that time spent on Census analysis will be profitable.

In addition to this work, the authors have developed a methodology that facilitates stratification of data into subsets that contain a simplified or single underlying function[14]. Results using property data[15] and Terrain data [16] indicate that this method significantly improves prediction accuracy.

The authors are currently in the process of combining the work presented in this paper with the stratification method[14], with the aim of grouping together enumeration districts that contain the same valuation functions. This will allow the modelling process to move from specialised models to locationally portable systems.

References

Millington A F, 1994, An Introduction to Property Valuation, Estates Gazette, 4th Edition.

Ashton P, 1972a, The Use of Multiple Regression Analysis in the Valuation of Real Estate, The Real Estate Appraiser,January 1972,pp 12-14.

Gronow SA, Ware JA, Jenkins DH, Lewis OM, Almond NI, 1996, A Comparative Study of Residential Valuation Techniques and the Development of a House Value Model and Estimation System. ROPA end of award report (Available as an occasional paper from University of Glamorgan).

Worzola E, Lenk M, Silva A,. 1995. An Exploration of Neural Networks and Its Application to Real Estate Valuation, Journal of Real Estate Research: 185-201.

Adair A S, and McGreal S, 1987, The Application of Multiple Regression Analysis in Property Valuation, Journal of Valuation, Vol. 6, pp 57-67.

Evans A, James H, Collins A, 1992, Artificial Neural Networks: an Application to Residential Valuation in the UK, Journal of Property Valuation and Investment, Vol. 11, pp 195-204.

Bruce RW, and Sundell DJ, 1977, Multiple Regression Analysis: History and Applications in the Appraisal Profession, The Society of Real Estate Appraisers, Vol. 43, pp 37-44.

Dale A, and Marsh C, 1993, The 1991 Census User's Guide, HMSO Publications.

CACI Information Services - ACORN™, CACI Ltd., e-mail: marketing @CACI.CO.UK.

Charlton M, Openshaw S, and Wymer C, 1985, Some New Classifications of Census Enumeration Districts in Britain: A Poor Man's ACORN, Journal of Economic and Social Measurement, Vol 13, pp 69-98.

Tazelaar J M, 1989, Neural Networks, BYTE, August p214.

DTI Guidelines for Neural Computing, DTI's Neural Computing Technology Transfer (NCTT) Programme.

Openshaw S, Wymer C, 1995, 'Classification and Regionalisation' in S.Openshaw (ed) Census Users' Handbook, Longman, London.

Lewis OM, Ware JA, 1997, A Novel Neural Network Technique for Modelling Data Containing Multiple Functions, Lecture Notes in Computer Science, Springer Verlag.

Lewis OM, Ware JA, Jenkins DH, 1997, A Novel Neural Network Technique for the Valuation of Residential Property, Journal of Neural Computing and Applications, Springer Verlag.

Ware JA, Lewis OM, Kidner DB, 1997, A Neural Network Approach to the Compression of Digital Elevation Models, 5th GISRUK Research Conference Leeds.

Appendix

Table 3 - Description of Data Extracted from Census Database.

Feature	Feature Composition	Parent Table
% Full Time Employed	{Total Employed Full Time/ Total Persons(TP)}*100	L08 - Economic Position: Residents aged 16 and over.
% Unemployed	{Total Unemployed/TP* 100}	L08
Ratio of People per Car	Total Persons/Total Cars	L21 - Car Availability
% Professional	100 * {Total Professional/ Total Economically Active(TEA)}	L91 - Social Class and Economic position: 16+
% Managerial and Technical	{Total Mang. & Tech/TEA}*100	L91
% Skilled Non Manual	{Tot Skilled Non Man/TEA} * 100	L91
% Skilled Manual	{Total Skilled Manual/TEA}*100	L91
% Unskilled	{Total Unskilled/TEA}*100	L91
% White	{Total White/TP} * 100	L09 - Ethnic Groups.
% Black Caribbean	{Total Black Caribbean/TP}*100	L09
% Black African	{Total Black African/TP}*100	L09
% Black Other	{Total Black Other/TP}*100	L09
% Indian	{Total Indian/TP}*100	L09
% Pakistan	{Total Pakistan/TP}*100	L09
% Bangladeshi	{Total Bangladeshi/TP}*100	L09
% Chinese	{Total Chinese/TP}*100	L09
% Asian	{Total Asian/TP}*100	L09
% Other	{Total Other/TP}*100	L09
% Born in Ireland	{Total Irish/TP}*100	L09
% Outright Owner Occupied	{Total Outright Owner Occupied/ Total Households (TH)}*100	L58 - Household Space Type; Tenure,Amenities
% Buying Owner Occupied	{Total Buying Owner Occupied/ TH}*100	L58
% Renting Privately Furnished	{Total Renting Privately Furnished/TH}*100	L58
% Renting Privately Unfurnished	{Total Renting Privately Unfurnished/TH}*100	L58
% Rented for Job or Business	{Total Rented for Job or Business/TH}*100	L58
% Rented Housing Association	{Total Rented from Housing Association/TH}*100	L58
% Rented Local Authority	{Total Rented from Local Authority/TH}*100	L58
% Detached	{Total Detached Properties/TH} * 100	L58
% Semi	{Total Semi detached Properties/ TH} *100	L58
% Terraced	{Total Terraced Properties/TH} * 100	L58
% Purpose Built Flat in Residential Building	{Total Purpose Built Flats in Residential Buildings/TH}*100	L58
% Exclusive Use of Amenities	{Total Exclusive use of bath or shower and inside WC/ TH} *100	L58
% No Central Heating	{Total No Central Heating/TH} * 100	L58

A6.4 D. H. Jenkins, O. M. Lewis, N. Almond, S.A. Gronow, and J. A. Ware

"Towards an Intelligent Residential Appraisal Model", Journal of Property Research, Spring 1999.

Towards An Intelligent Residential Appraisal Model

D. H. Jenkins, O. M. Lewis, N. Almond, S.A. Gronow, and J. A. Ware

University of Glamorgan, Treforest, Mid Glamorgan, UK.

Abstract

In the UK, and indeed in many countries, Direct Capital Comparison (DCC) remains central to the practice of residential property valuers. Theoretically well founded, statistically or heuristically based alternatives, usually embodying regression techniques, have failed to penetrate professional practice despite long pedigrees. For several years, neural networks, in which values are not so much derived or assigned but discovered, have also been propounded as potential alternatives. Yet DCC clings to its pre-eminent position because it is readily understood and it is thought to produce more accurate results.

In order to improve upon DCC, complementary or alternative methods will need to enhance accuracy, and be equally intelligible and transparent. This paper reports empirical findings, discusses some of the obstacles that will need to be overcome and some of the constituents that may comprise an improved model.

Keywords: Residential Property Appraisal; Direct Capital Comparison; Intelligent Systems; Professional Knowledge; Kohonen Feature Map

1. A complex problem

> "If you want to understand some aspect of the Universe, it helps if you simplify it as much as possible, and include only those properties and characteristics that are essential to understanding."
>
> Hari Seldon, "Prelude to Foundation", Isaac Asimov

Residential property valuation is often viewed as the Cinderella of professional appraisal practice. Comparatively little space has been devoted to the topic in the syllabi of UK higher education. Graduates, lured by the world of commercial property, aspire to apply "scientific" investment appraisal techniques. Researchers also see greater complexity in commercial markets (e.g. Connellan and James, 1996).

Such aspirations and views may be justified. Certainly, the markets for professional advice reflect or condition such views. Rewards for appraisal personnel are much higher in commercial than residential markets.

Nevertheless, residential appraisers are dealing with considerable complexity. First of all, housing is a complex commodity. At its most fundamental, the dwelling meets essential requirements for living: protection against climatic extremes, a place for activities preferably conducted under cover, a place for privacy and security, a place where physiological and psychological needs for territory are fulfilled. The dwelling will exhibit aspects of technological development in design and construction, often mediated through a regulatory framework. It will reflect communications technology in terms of its location and functionality. It will be an expression of the social needs and aspirations of individuals/ family units and the groups to which they belong or with which they identify. As such, dwellings are polymorphous and heterogeneous. In appraisal parlance, all houses are unique.

Next, housing markets are undeniably complex. In Marshallian economics the forces of supply and demand determine house prices. Rational people individually making decisions to maximise their utility in aggregate produce a competitive general equilibrium (in all markets simultaneously). In a perfect market resources are used efficiently and no one could be better off without violating this pareto-optimal allocation. This simplified model provides helpful insights into the derivation of value as Asimov predicts. In such a state, the value of a property newly arrived on the market could be deduced directly from analysis of the revealed prices (or rents) of comparable properties or equally directly computed from cost information.

It is safe to say, aside from Von Neuman's mathematical proofs that the conditions for a general equilibrium cannot exist (see Ormerod 1994), that none of the conditions of a perfect market begin to be met. The choice between value and cost information as the basis for a calculation becomes a pragmatic one: it concerns questions about the availability and reliability of data and its usefulness in the appraisal process, in predicting value.

Viewed from the level of the typical transaction, the price of a dwelling is an expression of its value to two individual household units at a point in time, the vendor and the purchaser and their agents. This price may also reflect the interests of third parties e.g. a moneylender or the taxman, the activities of other players in the market place whose signals have been interpreted by the parties, the agents acting for vendors of alternative properties, the unsuccessful bidders for the property sold and perceptions of general and local market conditions. The price will also reflect the state of other markets, markets for housing finance and finance more generally (assumed to be yet more unpredictable than housing markets (Mueller 1995)), and

substitute housing markets (non-owner-occupied housing) and factors which impact such markets e.g. forms of regulation and the degree of subsidy.

Furthermore, the value to the parties in a particular transaction may contain attributes to which every other purchaser may be indifferent. For example, environmental psychologists have analysed attachment to places, to neighbourhoods, towns and regions. Residential satisfaction is often tied to place attachment (Sundstrom et al 1996). What underpins this place attachment may not be clearly related to a measurable phenomenon of use in the comparison process, like proximity to a strong "attractor" e.g. school provision or pleasant environment, the substance of quantitative economics.

Aggregate house price functions have difficulty representing such imponderables, but the individual purchaser is prepared to pay a premium to realise the aspiration. It is a component of effective demand, mediated through the house purchase process. The purchaser's own "house price function" will reflect this factor, which may be measured by asking "How much are you prepared to pay to satisfy this want?"

What can be concluded from these observations?

Useful models for the estimation of house value are likely to be complex.

The analysis of one or more housing transactions is unlikely to be of more than indicative usefulness in the derivation of value of a neighbouring house.

The factors that determine house prices are not explained within any one discipline.

2. Dealing with complexity

> "Now, as you wish to know more and more about any phenomenon, or as a phenomenon becomes more complex, you need more and more elaborate equations, more and more detailed programming, and you end with a computerised simulation that is harder and harder to grasp."
>
> Hari Seldon, "Prelude to Foundation", Isaac Asimov

Ideally, the required output, a figure of value, would be solved by the use of a function or functions that could adequately handle such complexity, however hard to grasp (Asimov). The development of such a function is not imminent. The theoretical deduction of such a function has proven elusive and indeed Mason and Quigley suggest impossible (Mason and Quigley 1996). To date, there is not even a coherent classification of the inputs to the problem space - existing reductions do not include all "those properties and characteristics that are essential to understanding".

Where precise functions do not exist in situations exhibiting high complexity, heuristics are commonplace. Any heuristic that consistently provides tolerable results merits enduring use in the absence of better alternatives. However, if there is a lack of consistency or periodic failure, then improvements are desirable which attempt to identify weaknesses that cause inconsistency and/ or define the boundaries within which the heuristic may be usefully deployed.

The valuation of residential property for mortgage purposes is traditionally based on DCC, a heuristic which has an identifiable basis in known cognitive techniques. It has strength in that

Its theoretical precepts are readily assimilated. Houses may be compared by reference to their attributes, housing transactions by reference to market conditions. Valuers understand it.

It focuses the decision-maker on known quantities, earlier realised transaction prices revealed in the market. Valuers are able to identify inputs to it.

Its application is not demanding where data is adequate. Mackmin (1994 p45) defined the common processes in DCC (see below), though there is considerable scope for artful interpretation of method (see for example Dennett 1997). Valuers know how to use it.

However, DCC remains a heuristic. In relation to the complex model it seeks to interpret, its usefulness depends entirely on results. Gronow et al. (1996) establish that the heuristic broke down completely in the face of the broad market collapse of 1988/9 and that any usefulness took a considerable period to be re-established. Furthermore, various inherent weaknesses undermined real continuing usefulness while documented flaws in its application exacerbated this (ibid.).

While "a skilled valuer may use three or less comparables" in the analysis of any one transaction (Adair and McGreal, 1987), the typical mortgage valuer undertaking five appraisals per day relies upon substantial amounts of useful evidence. Yet transaction data in much of the UK remains confidential and there are no standards for its management. Adequate analysis requires sophisticated data manipulation, yet tools for the purpose remain undeveloped (Gronow et al. 1996). Revealed transaction prices contaminate the selection of evidence in the comparable process (Wolverton and Diaz 1996) and the valuation decision itself (Gronow et al. 1996). Yet this tentative transaction price is invariably revealed to valuers.

Finally, the valuation decision is shrouded in mystery. It is barely documented, as the courts have had cause to lament, and unexplained. Mackmin (1994) suggests that valuation by DCC can be broken down into four steps:

- Select comparables
- Extract, confirm and analyse comparable sale prices
- Adjust sale prices for noted differences
- *Formulate an opinion of open market value for subject property* (p 45)

Yet analysis of the valuation pro forma (those issued by lenders or designed by valuers) shows no space for any formal adjustment for noted differences or for exposing the rationale for the formulation of opinion. Observations of and discussions with valuers suggest that an attenuated form of this process occurs, though it is not committed to paper. The core of the decision process is akin to a black box.

However, during the entire period since the last market collapse, appraisal professionals in the UK nonetheless exclusively and continuously used DCC. The reason for this is straightforward. They perceive no viable alternative. Yet uncritical adherence to the heuristic implies that the profession is poised to repeat the mistakes of the last cycle.

The commonly used appraisal heuristic has often failed in its narrow purpose: the determination of house value for the purchaser. If the same heuristic also has a social function, the identification of inflationary and speculative tendencies in housing markets, it has failed in this broader purpose too. The social and economic consequences of a poor heuristic are evident. In the UK, alleged negligent valuations, in most cases calculated using DCC, accounted for over 30% of all cases listed in the Official Referee's court for action in 1996 (Mason and Rice, 1996). Loans to homeowners based on DCC calculations generated up to £10 billion of negative equity in the UK in this decade. Of course, the valuations were not the cause of negative equity, but they failed to apply any corrective and, in this sense, became a contributory factor (Jenkins, Rispin and Gronow 1995). The social consequences of mistaken valuations have been immense. Such consequences suggest a need for anything but complacency.

The remainder of this paper is concerned not with suggestions for the improvement of DCC techniques, which have been suggested elsewhere (Gronow et al., 1996). It is concerned with the elaboration of a model that would supplement the existing in the context of a problem statement (Section 3) and the report of empirical work designed to progress development of such a model (Section 4). This supplementary model will

need to be intelligible to its users. It will need some means of monitoring markets to identify discontinuities that catastrophically undermine the DCC heuristic. It will need an ample supply of data, not simply of the bricks and mortar type traditionally garnered by appraisers but also social and economic data that impact house prices. Further, it will need to embody fundamental domain knowledge gained from best practice that is appropriate across many appraisal paradigms. In short, to fulfil the role of satisfactory supplement, it will need to be an intelligent system.

3. Intelligent Systems

> "...with people and computers both on the job, computer error can be more quickly tracked down and corrected by people, and, conversely, human error can be more quickly corrected by computers."
>
> Hano Lindor "Prelude to Foundation", Isaac Asimov

Intelligent systems have been with us for some time. They have taken two principal forms. The more accessible form has represented decision processes in the form of rules. They have been labelled expert systems or expert database systems. A less accessible form has represented knowledge in the form of patterns. They have been labelled neural networks or genetic algorithms[6]. Case based systems might be understood as some form of combination of the two principal forms. However, case based systems are a special "case" rather than a possible synthesis of intelligent systems. Section 4 develops a notion of a hybrid in which an expert system acts as a front end to a hierarchy of neural networks. The remainder of this section considers some of the problems that need to be tackled and that have been highlighted as a result of current research.

3.1 Intelligible systems

Acceptability of any system requires that there be at least a modicum of understanding of the system by its operators. The appropriate degree of understanding is more closely analogous to the driving instructor than the mechanic, though some users will become mechanics. If an intelligent system is to be used by

[6] Genetic algorithms are inspired by the natural phenomenon of the 'survival of the fittest'. Problem solutions are evolved with each generation having a better solution than its predecessor (Goonatilake, et al, 1995).

appraisal professionals, they must have a reasonable understanding of its processes, be confident in explaining them and recognise when there is something wrong. At least, they must be able to explain decisions derived from system application to a standard equal at least to that thought apposite for DCC!

Expert systems are sufficient in this respect. A true expert system is always able to explain its decision process. Neural networks, however, are as much of a black box as DCC. To be acceptable, the neural network component of any intelligent system needs to be transparent. This is a challenge that cannot be avoided.

Understanding breeds confidence. Those who use the systems will need to be confident in their use. They need to know the limits to which the system may be safely driven. They will also need reassurance about system suitability when the terrain changes. For this reason, the intelligible system will also need to send messages to users about the degree of confidence that is appropriate to any current transaction. Intelligibility requires that even if the calculation of confidence is complex, utilising some variant of a conditional probability density function, nevertheless, the communication of confidence to the user is simple to understand.

3.2 Market Aware Systems

Cycles in UK residential property markets are quite well documented and a useful, general bibliography of property cycles has been produced by the Department of Land Economy at the University of Aberdeen (RICS, 1993). Since the early 1970's, there have been wide fluctuations in house prices from the equilibrium and a series of quite dramatic, asynchronous collapses. It is fair to say that few have predicted such collapses. Indeed it is noteworthy that the major players in housing markets - lenders, estate agents and, perhaps, valuers - retain an interest in promoting house price increases. Not only is professional fee income directly linked to transaction price, but also slightly rising markets have positive psychological effects: consumers are more confident and lenders' portfolios appear more secure.

Clearly, an intelligent system would require a time series analysis component. While precision in such predictions is unlikely and the estimation of turning points notoriously difficult, the identification of leading indicators to the appraiser may well serve to eliminate the wildest over/ under-valuations. Even if the basis of valuation were changed from Open Market Value to Estimated Realisation Price, or some other expression of a longer run equilibrium, a time series analysis component would be vital in identifying overly-strong bids.

The time series problem requires measurements of a single observable as a function of time as a basis for prediction. The first problem is selection of the "single observable". While this must be a measure of house price, there are many from which to choose, with different aggregates (national, regional and even sub-regional; by house type etc.) and periodic re-basing (e.g. Nationwide index in 1995). The second problem is the selection of measurement technique. Given the complexity of the market, it is assumed that in addition to linear and periodic components, there are random and chaotic components, the latter caused by non-linear dynamics. Because of this, our starting measure would be some form of feedforward neural network (Vermuri and Rogers 1994) though a comparison of neural network and earlier statistical methods may be helpful. A third problem is the period of time over which to perform the analysis.

3.3 Data Rich Systems

In addition to generalised cycles, there are myriad smaller relative price movements. These relate to the dynamics of location and perceptions of location at varying aggregates - region, sub-region, neighbourhood and immediate vicinity. They also relate to the dynamics of market sector, which has a complex interaction with location (e.g. Maclennan and Tu 1996) as well as strong independent defining features. Much locational and sectoral data needs to be recorded and monitored. As well as these more dynamic elements of the model there are the physical aspects of property - the bricks, mortar and topography captured in traditional approaches. Indeed, if the model is not to become saturated in supply side data, some measures of (a)typicality need to be developed to assist in modelling.

An intelligent system will draw on diverse data sources. Past developments of alternative approaches have stumbled for two reasons, first, because of a lack of data and poor data models. An intelligent system requires data models that are sufficient to represent underlying complexity, whether they are relational, object-oriented or some synthesis of these, while the system needs to remain sufficiently robust to provide estimations where there is little actual data.

Secondly, they have stumbled because of the elaborate nature of functions that are required to model this complexity without some form of decomposition. If the price or open market value of dwellings is determined by supply and demand, then value is discovered when simultaneous equations representing a function of demand and a

function of supply are solved. Given the multitude and complexity of factors which contribute to both the supply and demand functions, house-price determination studies, whether using neural networks, Multiple Regression Analysis (MRA), or other paradigms begin with a limited and manageable number of variables regarded as most sensitive. In this reduction, the open market value of the dwelling is associated, *ceteris paribus,* with the bundle of attributes it comprises, generally expressed in the following form:

$V = f(x, y, ... n)$

where the value V is some function of the attributes $x, y, ...n$.

In this particular reduction, demand related attributes, the most price-sensitive, are not explicit, there is no (or an isolated) supply effect over time and the properties are drawn from the same location or from locations which share identical features. As a result, MRA and neural network studies are constrained within narrow time and space zones by the strictures of the reduction (Gronow et al 1996). Recent neural network competitions organised by the International Association of Assessment Officers reflect this reality. Similar constraints have applied to expert system developments where multiple house price functions are heuristically assigned rather than statistically derived (Czernkowski, 1991; Jenkins and Gronow, 1993). The problem, in summary, is that *ceteris* is seldom *paribus*.

There are further problems with such approaches. It is established that property markets are differentiated spatially and sectorally so that choice of study area may help or hinder the explanatory power of the function. Different market segments are recognised by the presence of some attributes rather than others. Clearly, attributes typical of a particular segment need to be included within a function applied to that segment.

Nevertheless, such studies are potentially useful in assessing the relative value of properties within an economically static framework. The results may then be used, for example, for taxation purposes, though it is important to ensure that the model accommodates perceptions of taxpayers as to the constituents of value if the tax base is to be seen to be fairly assessed.

The limitations of such studies in estimating house prices in myriad small, dynamic market segments with poor information should be clear. Such models will need to

address the limitations imposed by the underlying assumptions before they can be used for the more problematic purpose of market appraisal.

If complex data models lead to over-elaborate functions - the number of variables and their interdependence creates more complexity than can be accommodated within existing models - then some preliminary classification of the input space is required. Attempts at classifying the variables are helpful therefore in that they may uncover a hierarchy of approaches to the problem and indeed may also identify surrogate measures that are capable of representing classes or sub-classes of attributes, thus simplifying functions. The valuers' mantra that value is dependent on location, location and location represents one well-known attempt at such a classification.

3.4 Knowledge Rich Systems

Eliciting knowledge from professional valuers is no small task. Domain knowledge is wide and deep, it is not always easily articulated (Scott 1988), conflict and contradiction are present and need to be managed (Nawawi et al 1996) and the interface between knowledge and data requires clearer elaboration (Jenkins 1992), especially in relation to local knowledge.

The representation of knowledge has been achieved for small scale, well-defined projects (Scott 1988, Jenkins 1992) using expert system shells and expert databases. The research and development effort required for modelling more complex knowledge is in its infancy.

The debate about the real content of professional knowledge is at least three cornered. Professional decision making may be a "conspiracy against the laity" or "problem solving made rigorous by the application of scientific theory and technique" or an intuitive and artistic process. The truth lies somewhere in the middle. A more definitive statement for the appraisal profession will require considerably more study than has occurred to date.

Scott (1988) and Jenkins (1992) noted in studies of appraisal decision-makers that though they were competent, they were nonetheless unable to articulate their full decision process. Schön (1991) suggested that this is a general phenomenon of professional decision-making:

> In his day-to-day practice he makes innumerable judgements of quality for which he cannot state adequate criteria, and he displays skills for which he cannot state the rules and procedures. Even when he makes conscious

> use of research-based theories and techniques, he is dependent on tacit recognitions, judgements, and skilful performances. (Page 50)

Even so he noted that this is not an unthinking approach:

> On the other hand, both ordinary people and professional practitioners often think about what they are doing, sometimes even while doing it. Stimulated by surprise, they turn thought back on action and on the knowing which is implicit in the action. (Page 50)

And he concluded:

> It is this entire process of reflection-in-action which is central to the "art" by which practitioners sometimes deal well with situations of uncertainty, instability, uniqueness and value conflict. (Page 50)

Schön described the content of this reflection-in-action as a form of experimentation, which, though less rigorous than controlled experiment, is nonetheless valid within the confines of the problem parameters. He talks about the development of a repertoire that is more subtle and flexible than the sum of case knowledge derived from practical experience.

His conclusion has three profound implications. The first concerns the way in which professionals are educated. This is not the immediate concern of this paper but see Schön (1987) for further discussion and an explanation of "the reflective practicum".

Secondly, there is an implication for our understanding of professional behaviour. For example, the observation that valuers are substantially influenced by revealed transaction price tells only part of a story. What surprises more is what valuers omit to consider and to analyse. Why do they not carry out a more rigorous analysis in keeping with the textbook? The answer would seem to be that once they have sufficient evidence to justify adoption of the tentative transaction price, they stop "experimenting". The answer may not be "correct" but it is sufficient within the parameters of the problem. These parameters include not only the technical, but also the contextual, the social, political and economic as well as the "lifeworld" and the personal

> This lifeworld is the complex array of routines, expectations, roles, norms, axioms and unwritten rules that make up the practitioner's everyday world. (James 1997 page 7).

In the instant example they include, inter alia, the size of the caseload, the amount of the fee, the lenders' need to secure the business, the practitioner's position in the company and the availability of evidence.

Thirdly there are implications for modelling the decision-making process. Formally, knowledge representation requires the faithful reproduction of the knowledge elicited - be it in the form of rules, procedures, cases or patterns. But caution should be exercised on two counts. First, when observed heuristics are reflecting a less than satisfactory treatment of a problem, as described in relation to DCC above. In such circumstances, the "knowledge engineer" needs freedom to incorporate the "deeper" knowledge, particularly where a more rigorous approach to the problem can be rendered to the system because the computer is not constrained in the same way as the human practitioner e.g. recursive trawls through data. Secondly, because the elicitation exercise might mimic the practice without the benefit of reflection on the reflection in action.

An intelligent system that is able to entertain the sort of dialogue with a given practice situation that results in a sufficient solution and self-learning appears remote. Schön's motivation may have been to develop an antithesis to the prevailing technical rational epistemology of the post-war period (Eraut 1995). The resulting synthesis, however, may be a more rigorous specification of a truly intelligent model. It should be clear by now that the development of a computer model to create a partnership with the human practitioner is not some ruse to supplant the practitioner but reflects, in Polanyi's phrase, a "tacit knowing" that the intuitive aspects of the decision-making process may defy systematisation. However, this does not prevent enquiry.

Mackmin (1994) identified the key stages in the valuation decision-making process. Gronow et al (1996) highlighted moments within these key stages where decision-making was blurred. Current research efforts are being focused on such moments and in particular on references to abstruse concepts like "local knowledge" to which an appeal is made when articulation runs dry (see Section 4.6).

4. Steps towards an Intelligent Hybrid

This section reports empirical work and preliminary findings from a programme of research geared towards the development of an intelligent system. The system is a hybrid in the sense, first, that it borrows from previous research into expert systems and neural networks and attempts to synthesise them and secondly in that it brings together traditional supply-side valuation knowledge with demand-side information.

4.1 Research objectives

One of the weaknesses faced in practice is the lack of comparable properties as evidence (Worzola et al. 1995). Given this weakness, much research has been carried out into directly calculating the value of a property from its locational and physical attributes (Borst 1991, Worzola et al. 1995, Adair and McGreal 1987, Evans et al. 1992, Do and Grudnitski 1992). Many of these research studies have considered the application of neural networks to residential property appraisal (Borst 1991, Evans et al. 1992, Do and Grudnitski 1992), with the majority of studies using data from a homogeneous area (i.e. an area where all properties are subject to the same environmental and locational forces) for the reasons explained. This approach is taken, as the valuation function can become very elaborate when spread across a heterogeneous area. This leads to neural network models that are not locationally portable. Nevertheless, the studies have reported a high level of success, with average absolute percentage error levels of between 5 and 7.5% not being uncommon (Evans et al.1992, Borst 1991).

The first objective of the research team was to break free from the limitations imposed by the study of a homogeneous area. If an intelligent system is to be developed, it must be capable of modelling a heterogeneous area without the need for constant recallibration. The novel approach taken is described in Section 4.2. The second objective was to represent within the functions developed some surrogates for demand. Indeed, it is submitted that the heterogeneity problem could not be fully solved without some reference to the relative wealth of market sectors and this is the subject of Section 4.3. Progress toward a third objective, the extent to which intelligibility may be built into a system, is described in Section 4.4. Barely yet addressed, the next objective is to address cycles in property markets. Section 4.5 sets out a perspective for addressing this "market awareness problem". The research team is concerned to harness more professional knowledge and Section 4.6 describes current research in this area. A summary of findings from empirical work to date and a schema of the intelligent model as it currently stands are to be found in Section 5.

4.2 The Heterogeneity Problem

In order to create a generic solution, it will be necessary to apply networks to a non-homogeneous area and extend the depth of attributes beyond traditional studies. The first step was to attempt to identify, from a heterogeneous parent dataset, a complete set of homogeneous sub-sets. The Cardiff area was chosen because access was available to a database that provided adequate records of an area that the research team considered to be heterogeneous in nature and of which they had some prior knowledge.

Adair et al. (1996) hypothesise that sub-markets can be identified by stratifying the market into increasingly homogeneous subsets. Using the hypothesis that a heterogeneous market consists of many homogeneous sub-markets, it is postulated that a heterogeneous market can be modelled indirectly using many models, trained on subsets of the parent dataset. To use such a system to predict the value of a previously unseen property then requires a method of selecting the appropriate Neural Network model. In this approach, a clustering algorithm was used both to identify groupings within the parent dataset and to act as a panel judge to decide which estimation model to select when asked to give a valuation.

The clustering algorithm used was the Kohonen Self Organising Map (Kohonen, 1984), which clusters similar input patterns. Figure 1 shows a typical 2D Kohonen Self-Organising Map along with an abridged algorithm (Note, the number of nodes are arbitrarily selected for example purposes).

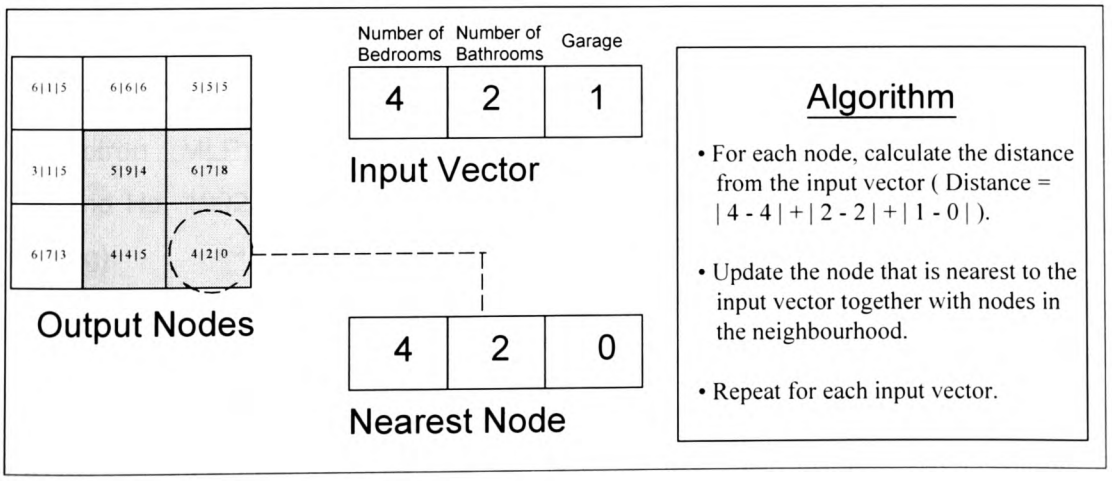


Figure 1 - A Kohonen Self Organising Feature Map.

Each output node on the Kohonen Feature Map contains a vector of length 'j', where 'j' is equal to the number of input attributes. Before training the network is in an initialised state (i.e. the directions of the vectors in each node are random). Training

involves passing an input vector into the network through the input nodes. Each node on the Kohonen Feature Map is then compared with the input vector, and the closest node is then changed to be more like the input vector. Neighbouring nodes also become more like the input vector. Iterating this process achieves clustering of similar input vectors in Euclidean space.

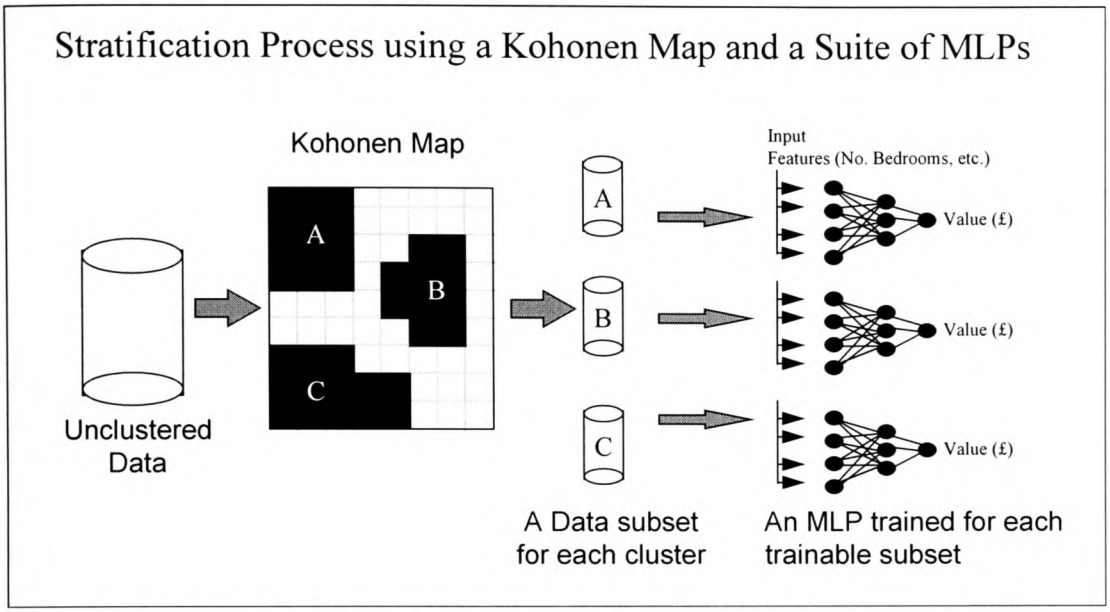


Figure 2 - An Overview of the Methodology. During Training, the whole historical dataset is separated - using a Kohonen Self Organising Map - into subsets that are subsequently used to train a series of multi-layered perceptron networks. During operation, the Kohonen Feature Map is used to determine which network to use to provide an estimate of value.

The methodology, illustrated in Figure 2 builds on research reported by James (1994). A Kohonen Feature Map is used to uncover sub-markets within a large dataset that are subsequently independently used to train a series of Multi-Layered Perceptron (MLP) networks trained using an error back propagation algorithm (see Tay and Ho, 1992 for an explanation of MLP networks and error back propagation training).

The advantage of using the Kohonen Feature Map for this application is that it can identify clusters within the parent dataset that are difficult to identify using simple sort procedures. However, it is sometimes difficult to identify class boundaries within a trained Kohonen Feature Map (James 1994) and this in turn leads to problems in generating training sets for the MLP networks. For example, consider the Kohonen Feature Map shown Figure 3; there appear to be five classes within the dataset, but there are regions of uncertainty relating to the boundaries of each cluster. The boundaries could be estimated visually, but at the expense of accuracy.

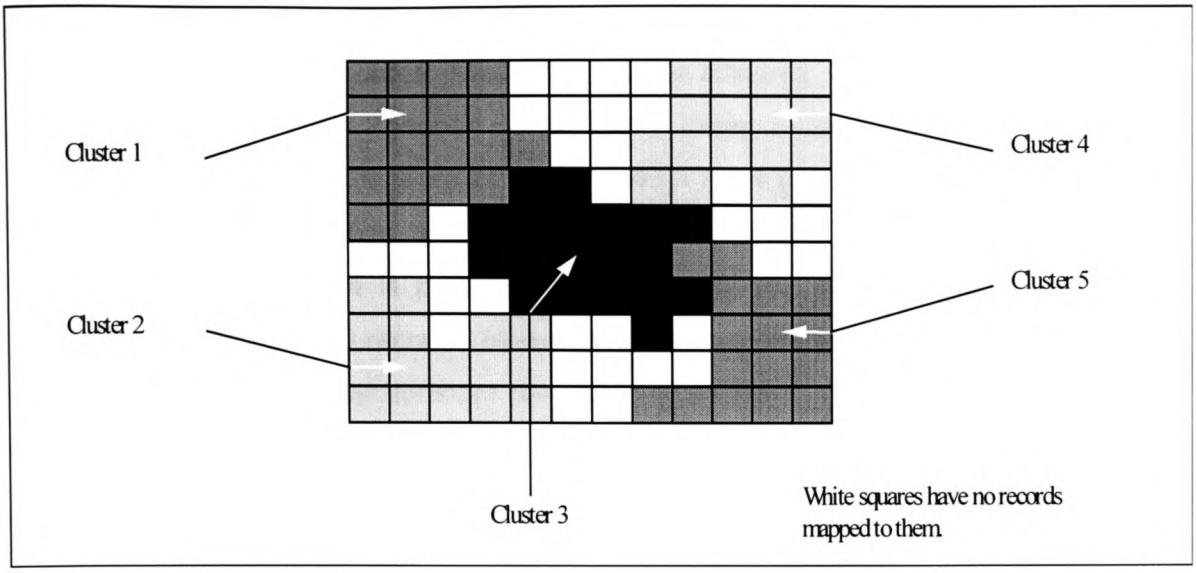


Figure 3 - An Example Trained Kohonen Self Organising Feature Map

Faced with this problem, in addition to standard techniques, it was decided to investigate a recently published variance estimation routine known as the Gamma test (Stefánsson, et al, 1997) with the expectation that this would enhance boundary detection (Lewis, et al, 1997a, Lewis, et al, 1997b).

A database containing information on residential property transactions during the period January 1993 to December 1995 was selected to test the methodology. There are 990 records in the original database, each with 51 attributes. However, a number of these have either constant values or free form text that is difficult to recode and were therefore removed. A description of the database used for this study is shown in Table 1.

Table 1 - A description of the database.

Attribute Name	Example Value
Street Name	Newport Road
District or Village	Roath
Unit	1 - 6
Unit Type	Mid terraced etc.
Unit Size	Area M^2
Attribute Name	Example Value
Valuation Date	19 May 1995
Main Heating	Full, Partial, None
Number of Bedrooms	1 - 8
Age in Years	0 - 500
Number of Garages	0 - 2
Value	10,000-255,000

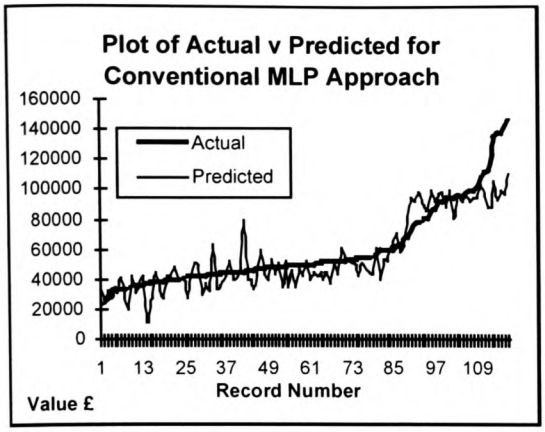
A 10 by 10 Kohonen Feature Map was used to find the groupings in the historical dataset containing 990 records. Value, the output attribute, was omitted from the Data Set used to train the Kohonen Feature Map. Using a combination of boundary detection methods, the data were found to contain eight groups. The records from each group were examined and common attributes within the group removed. Obviously, as the dataset is partitioned into classes, the classes contain only a portion of the original 990 records. However, this is accompanied by a decrease in the number of attributes - as constant columns are removed.

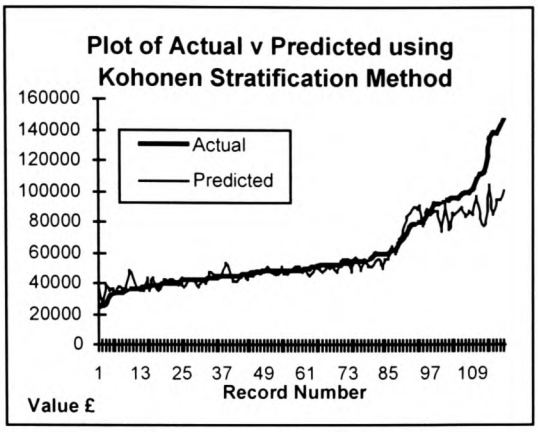
In order to provide a benchmark for analysing the methodology, a single MLP model and a single MRA model were constructed using the whole dataset. After training the ability of the models to appraise residential properties with known values was tested. The MLP model outperformed the MRA model and hence the results of the MLP model shown in Figure 4a are used as a benchmark; the graph shows the actual and predicted values for 117 test properties. Table 2 illustrates the results achieved using the described methodology (Kohonen stratification model) on the same 117 properties in the test set.

Table 2 - Results achieved for the test set.

	Conventional MLP	**MRA**	**Kohonen Stratification**
Mean absolute % error	18%	24%	8%
% of Records with an error > 10%	74%	79%	22%
Minimum absolute % error	0%	0%	0%
Maximum absolute % error	310%	220%	49%

Figure 4b shows the improvement in accuracy using the new method over the conventional MLP approach (the data used for Figure 4a and Figure 4b have been sorted in ascending actual property value order).





(a) *(b)*

Figure 4

(a) A graph of actual and predicted value gained using a conventional MLP approach.

(b) A graph of actual and predicted value gained using the Kohonen stratification methodology.

It is evident, from the results obtained that the methodology compares very favourably with the more conventional neural network approach and the multiple regression approach. An average increase in prediction accuracy of 10% was achieved using the new method over the conventional approach. This implies that the original dataset either contained more than one underlying function (James 1994) or the function was too elaborate to be modelled using a single MLP network. Moreover the Kohonen Feature Map can discern different classes within the data, which when independently modelled yield a greater predictive accuracy than those computed for the original dataset (James 1994). In this way, the system may transcend the limitations implied by traditional locational versus sectoral searches. Sufficiently refined clusters may represent specific property types in specific sub-markets.

However, further analysis suggested that clusters formed by the Kohonen Self Organising Map did not always lead to dramatic increases in prediction accuracy. This was to be expected, as no real indicators of location were included in the analysis. To address this problem, "demand-side data" as described in Section 4.3 was considered.

4.3 The Demand-Side Problem

In the first approach described in Section 4.2, two numeric surrogates (Street and District) were used to represent location, and indirectly wealth. This was without

doubt a poor choice. An alternative method of ranking wealth is to consider average house value in each district. Jenkins (1992) used this method to generate "base values" in a heuristically based valuation system. In a second approach, the average house value for each district was used as an input to the MLP model.

From the results obtained, it was clear that an increase in modelling accuracy could be gained. In fact, the best results were achieved by finding average values for each house type in each district. This avenue of research will no doubt be revisited in the near future. However, it is imperative that more depth is added to the database from other sources as the variance within the dataset was still outside any acceptable threshold. In order to improve modelling accuracy further, the use of Census data was investigated.

Census data is available at a number of abstraction levels with the smallest being Enumeration District (ED) in England and Wales, and Output Area (OA) in Scotland. Originally EDs were intended to contain between 15 and 200 inhabited houses, with current ED sizes representing a workload that can be performed by a single enumerator in the time available, given the circumstances of the area. In the 1991 Census there were 106,866 ED's in England and 6,330 in Wales. A coding system allows access keys to be generated by following a country/county/district/ward/ED route. In addition to this a cross-reference from Postcode to ED allows individual properties to be included in ED level statistics (Dale and Marsh, 1993).

Ideally, to achieve maximum benefit from the Census data, the Cardiff dataset should be expanded at Postcode/ED level. As this is labour intensive, it is important that only useful information is extracted from the Census database. The empirical results described in this section used the Census data described in Table 3.

Table 3 - Census Variables Used in Analysis

Socio-Economic Group	
Employers and Managers (Large est.)	Employers and Managers (small est.)
Professional workers (self-employed)	Professional workers (employees)
Ancillary workers and Artists	Foreman and Supervisors (non-manual)
Junior non-manual workers	Personal Services workers
Foreman and Supervisors (manual)	Skilled Manual workers
Semi-Skilled Manual workers	Unskilled Manual workers
Members of Armed Forces	
Employment	
Full-Time Employment	On Government Scheme
Part-Time Employment	Unemployed
Self Employed	
Qualifications	
Qualified Persons	Higher Degree
Degree	Diploma
Qualified and on Government Scheme	Qualified and Unemployed

Age Ranges of Qualified Persons	
Housing Stock	
Detached Properties	Purpose-Built Flats
Semi-Detached Properties	Converted Flats
Terraced Properties	Bedsits
Tenure	
Owner Occupied (Outright)	Owner Occupied (Buying)
Privately Rented (Furnished)	Privately Rented (Unfurnished)
Rented from Housing Association	Rented from Local Authority
Amenities	
Shared Use of WC	Exclusive Use of WC
Central Heating	
Availability of a Car	
Households with no car	Households with 1 car
Households with 2 cars	Households with 3+ cars
Ethnicity	
White	Black Caribbean
Black African	Black Other
Indian	Pakistani
Bangladeshi	Chinese
Asian	Persons born in Ireland
Miscellaneous Variables	
Working Mothers (Part-Time)	Working Mothers (Full-Time)
Lifestages (age ranges of residents)	Overcrowding (persons per household)
Travel to work estimates	

The initial research, described more fully elsewhere (Lewis, et al, 1997c), linked the same house data with Census[7] data at the District and ED level. The Cardiff data set was reduced to 660 records for the ED level analysis due to missing or incomplete values in the postcode attribute. Some improvement in accuracy was gained even at the very highest level of abstraction. As anticipated, however, the greatest gains occurred at the ED level, where the Mean Absolute % Error was reduced from 20% (original Cardiff Data Set) to 13% (Cardiff Dataset & ED level Census statistics).

Introducing Census data at the ED level clearly improved the prediction accuracy of the single MLP network. Having constructed a single MLP model using Census data, the next logical step was to include Census data in the Kohonen/MLP methodology previously described. Figure 5 presents a framework for including Census data in the stratification model.

[7] "Source: The 1991 Census, Crown Copyright. ESRC purchase".

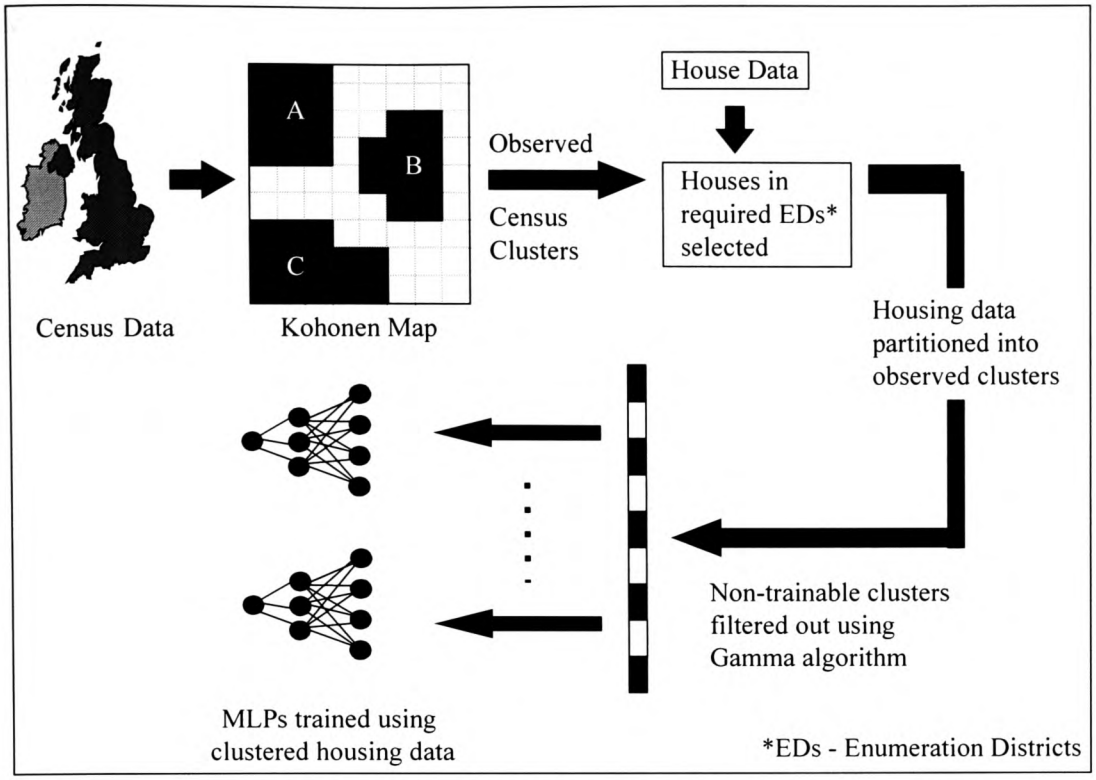


Figure 5 - A Framework for Including Census Data in the Stratification Model.

In order to test the effectiveness of the method for determining useful residential property sub-markets, two testing procedures were developed. The first concerned the ability of the sub-models to outperform a single model trained on all of the domain data. Sub-models were constructed based on individual geodemographic factors, for example *housing stock in region* and *employment statistics.* Test properties were passed through the single MLP model trained on all of the available data and also through the appropriate sub-models. An accuracy increase of between 1% and 14% was achieved by the sub-models compared with the single model predictions.

Table 4 - Sample Results for Two-County Analysis

Data Sample	Mean Absolute % Error	Improvement
County_B Whole Data Set Model	18%	-
County_A Model_A	18%	0%
County_A Model_B	17%	1%
County_A Model_C	15%	3%
County_A Model_D	15%	3%
County_A Model_E	18%	0%

The second testing procedure aimed to determine whether a sub-model trained using data selected from one geographical area could effectively be used to predict the

values of properties from a different geographical area sharing similar Census characteristics. MLP models were created using property data selected from one county in South Wales (County_A) and were used to predict the values of residential properties in a neighbouring county (County_B). The results (see Table 4) show that the County_A models at least match, and in some cases outperform, the predictions made by the single MLP model trained on County_B properties. Clearly, the ability of these models to predict the values of residential properties outside the geographical area from which the training data was selected has been demonstrated.

4.4 The Intelligibility Problem

One of the major criticisms of neural networks that discourage acceptance by professionals, including residential valuers, is their inherent 'black box' nature. The internal workings of a neural network are, for the most part, hidden from the user. For a residential valuer, this leads to a situation where one problem is solved only for another to be created; a house may be valued accurately using neural network techniques but there is no explanation as to how the value was deduced. One of the aims in building an intelligent system is to make neural network models more 'human-readable'.

> "The conventional approach to building an expert system requires a human expert to formulate the rules by which the data can be analysed" (Zurada, 1992, p. 225).

In contrast, a connectionist or induction expert system formulates its knowledge base by modelling implicit functions or relationships within a dataset. A connectionist expert system is in essence a straightforward neural network. However, the environment software examines the weights recorded for a trained network and attempts to give a tentative explanation of the model based on the magnitude of the weights. Unfortunately, the explanations generated by a connectionist expert system are far removed from the symbolic heuristics taken by a human expert and more in tune with the synoptic activity of the brain. Mozer and Smolensky suggest that the *"one thing that connectionist networks have in common with brains is that if you open them up and peer inside, all you can see is a big pile of goo"* (Mozer & Smolensky, 1989). However, despite this complexity, many researchers are investigating methods of extracting rules from trained networks and this route will undoubtedly be

considered by the Research Team in the near future (see Andrews, et al, 1995 for an introduction to rule extraction techniques).

Given the problems and inherent complexity associated with direct rule extraction from neural network structure it was decided to identify rules that describe the decisions made by the Kohonen Feature Map as opposed to the

IF Property_Type is "converted" AND Central_Heating is "none" THEN Network = "NET_A"	IF Building_Type is "house" AND Property_Type is "detached" AND Central_Heating is "full" THEN Network = "NET_B"
IF Building_Type is "house" AND Central_Heating is "partial" THEN Network = "NET_C"	IF Building_Type is "house" AND Property_Type is "mid terraced" AND Central_Heating is "full" THEN Network = "NET_D"
IF Building_Type is "house" AND Property_Type is "detached" AND Central_Heating is "full" AND Has_Garage is "no" THEN Network = "NET_E"	IF Building_Type is "house" AND Property_Type is "end terraced" AND Central_Heating is "none" THEN Network = "NET_F"
IF Building_Type is "house" AND Property_Type is "semi-detached" AND Central_Heating is "none" THEN Network = "NET_G"	IF Building_Type is "house" AND Property_Type is "semi-detached" AND Central_Heating is "full" THEN Network = "NET_H"

Figure 6 - A simple Rule Base for Model Selection

predictions made by the MLP networks. By examining the way the modular system works, it is evident that each sub-model (which in essence represent a homogeneous dataset) is constructed using a subset of the attributes present in the parent dataset. Moreover, this subset of attributes differs from one sub-model to the next, therefore, rules can be extracted (by inspection of the attributes subset) that describe which sub-model can best predict the value of a previously unseen property. An example rule base (shown in Figure 6) has been constructed for the supply side analysis previously described.

The rule base shown is somewhat crude and in no way describes the full complexity of either the data, or the appraisal process. It is obvious to anyone familiar with the valuation process that high level heuristics should contain reference to location. However, the inadequacy of the rule base is merely a factor of the quality and representation of the data. Employing a similar technique for the analysis including Census data reveals much more. Figure 7 gives some examples of the rules obtained.

Rule Extraction Results for Cluster 4 in Houstype Data

RULEBASE

IF [%of DETACHED properties in surrounding area] is in range 0 to 52
AND an average value of 21
AND a median value of 21
AND 1st and 3rd Quartile values of 3 and 35

AND [%of SEMI-DETACHED properties in surrounding area] is in range 0 to 48
AND an average value of 21
AND a median value of 21
AND 1st and 3rd Quartile values of 12 and 29

AND [%of TERRACED properties in surrounding area] is in range 0 to 60
AND an average value of 35
AND a median value of 37
AND 1st and 3rd Quartile values of 22 and 49

AND [%of PURPOSE-BUILT FLATS in surrounding area] is in range 0 to 184
AND an average value of 27
AND a median value of 20
AND 1st and 3rd Quartile values of 4 and 40

AND [%of CONVERTED FLATS in surrounding area] is in range 0 to 15
AND an average value of 1
AND a median value of 0
AND 1st and 3rd Quartile values of 0 and 1

AND [%of BEDSITS in surrounding area] is in range 0 to 0
AND an average value of 0
AND a median value of 0
AND 1st and 3rd Quartile values of 0 and 0

THEN prediction should be made using MLP Network: HTYPE_4

Rule Extraction Results for Cluster 4 in Car Data

RULEBASE

IF [Percentage Households with No Car] is in range 12 to 40
AND an average value of 30
AND a median value of 31
AND 1st and 3rd Quartile values of 26 and 35

AND [Percentage Households with 2 Cars] is in range 0 to 23
AND an average value of 15
AND a median value of 15
AND 1st and 3rd Quartile values of 12 and 17

AND [Percentgae Households with 3+ Cars] is in range 0 to 11
AND an average value of 3
AND a median value of 2
AND 1st and 3rd Quartile values of 2 and 4

AND [Percentage Households with 1 Car] is in range 46 to 69
AND an average value of 52
AND a median value of 52
AND 1st and 3rd Quartile values of 49 and 55

THEN prediction should be made using MLP Network: CARS_4

Figure 7 - Examples of Rules Extracted from Kohonen Self-Organising Map

The rules describe the profile of an area in terms of selected geodemographic characteristics. Quartiles are used to give an indication of the core of the profile ignoring tail ends. Each Quartile contains 25% of the distribution with Q1 and Q4 being the lower and upper tails respectively.

Although perhaps naive, these rules have obvious links with heuristics used as part of the appraisal process. As such the potential for 'human-readable' neural network modelling is modestly exhibited. The rulebase is in effect providing the reasoning for the network decision.

Given the complexity inherent in the problem, such reasoning, even after considerable training, is unlikely to be infallible. However, it may be synthesised with the reasoning used by human experts obtained through knowledge elicitation. In this way, an intelligent front end to the neural networks can be developed.

4.5 The Market Awareness Problem

In an earlier section the use of average house value was described as a crude means of representing wealth in the analysis. Apart from other limitations, in the context of residential appraisal, average house value is a retrospective measure of wealth. In order to address the market awareness problem, the system requires some means of anticipating where a market is going, of locating its position in a cycle. To achieve this, it would seem that a synthesis is required of research achievements at the micro-level to forecast the values of a single commercial property or small portfolio (Connellan and James, 1996) with research achievements at the macro-level in the modelling of the UK housing market (Miles and Andrew, 1997). In singling out such quantitative approaches for further elaboration, it is not intended that qualitative approaches to forecasting, like analogue forecasting or the Delphi technique, be disregarded. These are wholly appropriate in the development of valuation technique, especially as regards human processes. Rather, the intention is to focus on the aspects of valuation that are most susceptible to the development of computer modelling.

The first task involves identifying (a) leading indicator(s) that a system then tracks. Such an indicator might be a composite of well-recognised broadly based factors, disposable income, borrowing rate, inflation rate, and more (sub-) market particular factors, such as local employment and investment trends, which are part and parcel of an elusive component of the valuation paradigm known as "local knowledge". This is a relatively straightforward task and elements of it were accomplished in the development of an earlier prototype system (Jenkins et al 1995). Current research is engaging the issue of "local knowledge" in order to elaborate the earlier prototype.

The second task involves the utilisation of such indicative data in forecasting, in what Connellan and James (1996) have termed longitudinal analysis as opposed to the more traditional cross-sectional form. This requires confronting the problems identified in Section 3.2 above and is the penultimate stage in the current research agenda.

4.6 The Problem of Professional Knowledge

The research team is concerned with two overlapping tasks in the pursuit of professional knowledge. The first task is to gain a closer understanding of "local

knowledge". Practitioners in the UK are regulated by the RICS in terms of the geographical area in which they are qualified to practise. Competence is associated with local knowledge (though see Abbey National Mortgages plc v Key Surveyors Nationwide Ltd. And others, [1996] 33 EG 88). The first attempt at knowledge elicitation (Scott 1988) focused on general declarative knowledge with the aim of developing a rule-based expert system that was capable of undertaking mortgage valuations. The system built had severe limitations and there was considerable degradation in valuation results when the system was transported even within a region. This was partly addressed in subsequent research (Jenkins 1992) when valuers were encouraged to heuristically assign values to localities. However, the latter project made no attempt to classify local knowledge and the system was not transportable by design. The objective of current research is to resolve the paradox of local knowledge and transportable expertise so that such knowledge can be represented to the intelligent system.

A current experiment is attempting to establish whether local knowledge exists and to discover its content. In the experiment, local and "visiting" valuers are providing appraisals of a limited number of properties. Each is in receipt of the same data about the property and about comparable properties. The assumption is made that there are no unrecognised defects to the property. Each valuer is completing the same lender's form and is providing any notes of their appraisal decision and a log of time spent in the process. Care is being taken to ensure that valuers with similar experience and qualification are selected in order to isolate the local knowledge component.

The second stage of the research concentrates on in-depth interviews to compare the conceptions of local knowledge that valuers have, in part by focusing on the advantages/ disadvantages that participating valuers experienced.

The second task in relation to professional knowledge is an attempt to reconcile the heuristics from the local knowledge elicitation process with the rules extracted from the neural network approach. This is necessarily the final stage of the research program intended for reporting in 1999.

5 Conclusions

5.1 Summary of Research Findings

The empirical evidence leads to the following conclusions:

- A set of models, each dedicated to a certain narrow domain, can significantly outperform predictions made by a single more general model trained on all of the available training data.
- Demand side data, like Census data, enhances the efficacy of the model. The geographical size of the Census area selected affects its performance as a locational surrogate in an appraisal model. The advised aggregate to use to represent location is the Enumeration District.
- Many different features describe the characteristics of an area. These include obvious features such as 'housing stock in neighbouring region', but also include less obvious features such as 'number of cars per household'.
- Models created from the stratification technique can be used to predict property values in other areas that have similar Census characteristics.
- Using a Kohonen map to stratify the Census data, and representing the trained Kohonen map using simple statistics, provides additional transparency to the intelligent appraisal model.

5.2 Model Development

Having performed this analysis, it is now believed that further improvements can be made to this part of the model by adding another layer that automatically selects the most appropriate Census data to segment the property market. This step is currently done manually, though the Research Team intend investigating tree induction techniques (see Quinlan, 1986) and genetic algorithms (see Goldberg, 1989) to automate this step and ensure inclusion of influential variables for all models. Given that Census data degrades in quality between successive Censuses, alternative data sources will also be reviewed.

At the halfway point in the research programme, Figure 9 provides a schema of the hypothesised intelligent model. Inputs to the processes described can be undertaken independently. In the valuation process, the user interface requests data from the user and automatically selects a valuation function identified from the inputs.

This function is one of the many such functions produced by the training process described in Figure 5. The function applied to variables from the subject property computes the raw valuation figure. The raw figure is supplied to a second process that adjusts it to reflect changes in demand side trend.

Alongside the raw figure of value is a tolerance level for the prediction together, ultimately, with an explanation of the network's reasoning.

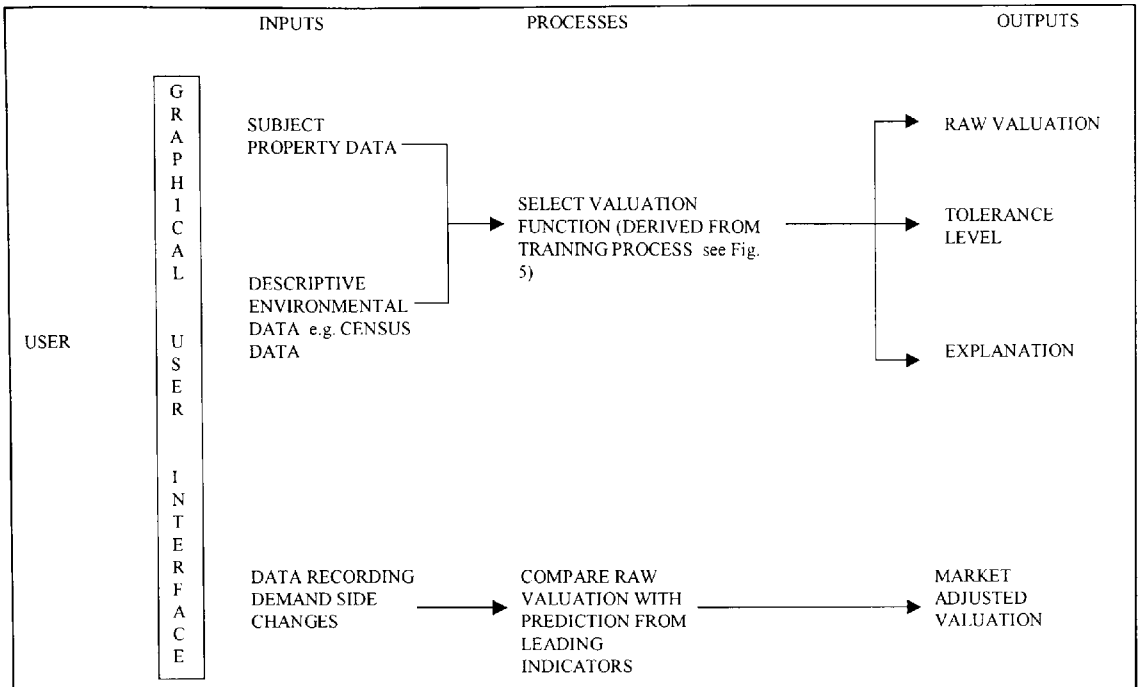


Figure 8. The "raw" valuation output of the valuation process becomes an input to the "market awareness" process.

5.3 Epilogue

Our first working hypothesis was that useful models for the estimation of house value are likely to be complex. This has been confirmed by experience. The development of computer based models to date seems to have been the preserve of academia. This should not be surprising given the artificial but real division between theory (universities) and practice (professional practitioners). Perhaps just as unsurprising (for the same reasons) practitioners have been reluctant to embrace such models despite the partnership nature of the proposals. If practitioners are unable to trust computer-based partners, perhaps the time is right for some friendly competition to demonstrate their usefulness.

While the intelligent system described here in outline remains a little way off (the research has not even begun to explore the potential contribution from other disciplines) the cumulative impact of the developments described above promises to increase the appraisal capacity of existing systems substantially.

It is proposed that a series of appraisals are conducted "in the laboratory" to compare the results obtained by a professional valuer using the traditional technique (though without prior knowledge of the tentative transaction price) and the far-from-intelligent

systems built to date. The objective is to demonstrate the usefulness of intelligent systems to humans and not the other way around - yet.

6. References

Adair, AS, Berry, JN, McGreal, WS, 1996, "Hedonic Modelling, Housing Submarkets and Residential Valuation", Journal of Property Research, Vol. 13, 67-83.

Adair, AS, and McGreal, S 1987, "The Application of Multiple Regression Analysis in Property Valuation", Journal of Valuation, Vol. 6, 57-67.

Andrews, R, Cable, R, Diederich, J, Geva, S, Golea, M, Hayward, R, Ho-Stuart, C, Tickle, A B, 1995, "An Evaluation and Comparison of Techniques for Extracting and Refining Rules from Artificial Neural Networks, in Knowledge-Based Systems Journal Vol 8, No 6 (December 1995).

Borst, RA 1991, "Artificial Neural Networks: The Next Modelling/Calibration Technology for the Assessment Community ?", Journal of Property Tax : 10,1,69-94.

Czernkowski, R, 1990, Expert Systems in Real Estate Valuation, Journal of Valuation, Vol. 8(4), pp 376-393.

Connellan, OP, and James, H, 1996, "Estimated Realisation Price by Neural Networks", RICS, Cutting Edge

Dale, A, and Marsh, C, 1993, "The 1991 Census User's Guide", HMSO Publications.

Dennett, RM, 1997, "The development of a fully integrated information technical solution to the residential property valuation process." Unpublished MPhil Thesis.

Do, Q, & Grudnitski, G, 1992. "A Neural Network Approach to Residential Property Appraisal", The Real Estate Appraiser, 38-45.

Eraut, M, 1995, Schön Shock: a Case for Reframing Reflection-in-Action?, *Teachers and Teaching: Theory and Practice*, Vol. 1(1), pp 9-22.

Evans, A, James, H, and Collins, A, 1992, "Artificial Neural Networks: an Application to Residential Valuation in the UK", Journal of Property Valuation & Investment: 11,195-204.

Goldberg, D E, 1989, Genetic Algorithms in Search, Optimization and Machine Learning, Adison-Wesley Publishing Company Inc. ISBN 0-201-15767-5

Goonatilake, S, & Khebbal, S, 1995, Intelligent Hybrid Systems, John Wiley & Sons.

Gronow SA, Ware JA, Jenkins DH, Lewis OM, Almond NI 1996, "A Comparative Study of Residential Valuation Techniques and the Development of a House Value Model and Estimation System", ESRC ROPA Report.

James C 1997, "How do you do? An introduction to Professional knowledge and its Development", University of Glamorgan

James, H, 1994, "An 'Automatic Pilot' for Surveyors", RICS Cutting Edge Conference

Jenkins, DH, 1992, "Expert Systems in the Land Strategy of Cardiff City Council", M Phil Thesis

Jenkins, DH, and Gronow, S, Expert Systems Shells: A Retrospective, Journal of Property Valuation and Investment Vol. 11:3 1993

Jenkins, DH, Rispin, CRW, and Gronow, S, 1995, "An integrated residential valuation system", First International Real Estate Society Conference, Stockholm.

Kohonen, T, (1984), A Simple Paradigm for the Self-Organised Formation of Structured Feature Maps, in Competition and Co-operation in Neural Networks. ed. S. Amari, M. Arbib. vol. 45. Berlin: Springer Verlag.

Lewis, O M & Ware J A, 1997a, A Novel Neural Network Technique for Modelling Data Containing Multiple Functions, in Computational Intelligence - Theory and Applications, ed. Bernd Reusch, (Lecture Notes for Computer Science Series Vol. 1226), Springer. ISBN 3-540-62868-1, pp 141-149.

Lewis OM, Ware A, and Jenkins DH, 1997b, "A Novel Neural Network Technique for the Valuation of Residential Property", Journal of Neural Computing and Applications, Springer Verlag.

Lewis, OM, Ware, JA and Jenkins DH 1997c, "The Use of Census Data in The Appraisal of Residential Properties Within the United Kingdom: a Neural Network Approach", 5th European Conference and Exhibition on Geographical Information Systems, Vienna

Mackmin, D, 1994, "The Valuation and Sale of Residential Property", Routledge, 2nd Edition.

Maclennan, D and Tu, Y, 1996, Economic perspectives on the structure of local housing systems, Housing Studies, Vol. 11(3), pp 387-406

Mason, C, and Quigley, JM, 1996, "Non-parametric hedonic housing prices", Housing Studies Vol. 11(3), pp 373-385.

Mason, J ,and Rice, R, 1996, "Property valuers win landmark ruling", Financial Times 21/6/96, p 8.

Miles and Andrew, 1997, "The Merrill Lynch Model of the UK Housing Market".

Mueller, GR, 1995, "Understanding real estate's physical and financial market cycles", Real Estate Finance 12 (3)

Mozer, M C, & Smolensky, P, 1989, "Using Relevance to Reduce Network Size Automatically", Connection Science, 1, 3-16.

Nawawi, AH, Jenkins, DH, and Gronow, SA, 1996 "Computer Assisted Rating Valuation of Commercial and Industrial Properties in Malaysia: Developing an Expert System from a Multiple Experts Knowledge Elicitation Methodology", RICS, Cutting Edge

Ormerod, P, 1994, "The Death of Economics," Faber & Faber.

Quinlan, JR, 1986, Induction of Decision Trees, Machine Learning, Vol 1, pp 81-106.

RICS 1993, Understanding the Property Cycle, Working Paper Two: A Literature review

Schön, DA, 1991, "The Reflective Practitioner", Basic Books Inc., 2nd Edition

Schön, DA, 1987, "Educating the Reflective Practitioner", Jossey-Bass Inc.

Scott, IP, 1988, "A Knowledge Based Approach to the Computer-Assisted Mortgage Valuation of Residential Property", PhD Thesis

Stefánsson, A, Koncar, N, Jones, A T, 1997, "A Note on the Gamma Test", Journal of Neural Computing and Applications, Vol. 5 Number 3, Springer Verlag.

Sundstrom, E, Bell, PA, Busby, PL and Asmus, C, 1996, Environmental psychology, Annual Review Psychology, 47, pp 485-512.

Tay, D P H, & Ho, D K H, 1992, Intelligent Mass Appraisal, Journal of Property Tax Assessment and Administration, Vol. 10, pp 5-25.

Vermuri and Rogers 1994, Artificial Neural Networks Forecasting Time Series, IEEE Computer Society Press, California

Wolverton, M and Diaz, J, 1996, Investigation Into Price Knowledge Induced Comparable Sale Selection Bias, RICS Cutting Edge Conference, Bristol.

Worzola, E, Lenk, M, Silva, A, 1995, "An Exploration of Neural Networks and Its Application to Real Estate Valuation", Journal of Real Estate Research: 185,201.

Zurada, JM, 1992, "Introduction to Artificial Neural Systems", West Publishing Company (ISBN 0-314-93391-3) p58.

A6.5 Almond, N.I., Lewis, O.M., Jenkins, D.H., Gronow, S.A. and Ware, J.A.

"Intelligent Systems for the Valuation of Residential Property", Royal Institute of Chartered Surveyors Cutting Edge Conference, 1997.

INTELLIGENT SYSTEMS FOR THE VALUATION OF RESIDENTIAL PROPERTY

Nigel Almond*, Owen Lewis, David Jenkins
Stuart Gronow and Andrew Ware

***Contact Author**

Centre for Research in the Built Environment
University of Glamorgan
Pontypridd, Mid Glamorgan, CF37 1DL, UK
tel. +44 (0) 1443 482708
fax. +44 (0) 1443 482660
e-mail*. nalmond1@glam.ac.uk

ABSTRACT

In the majority of cases, the valuation of residential property for mortgage purposes within the UK is performed using the method of Direct Capital Comparison (DCC), a process which, appears to be widely understood, correct and sufficient given the limited space devoted to the subject within appraisal texts. However, the collapse of the UK housing market in the late 1980's and the consideration of alternative approaches to valuation has fuelled concern as to the applicability of the method as currently applied in practice to the task in hand.

Research at the University of Glamorgan, more recently funded through a Realising Our Potential Award from the ESRC has highlighted a number of deficiencies in the approach, not only in its application by practitioners, but also in the context of the supply of valuation advice. Drawing upon this research a number of potential solutions are discussed.

The paper also reports on discussions among behavioural psychologists about problems associated with professional knowledge and considers the ramifications of these discussions for valuation practice and education.

1. INTRODUCTION

The valuation of residential property for mortgage purposes can conceivably be performed using one of four methods, i.e. the comparative, income, cost or residual approach. However,

in practice, it is the comparative method, often referred to as DCC that is most widely used in determining the Open Market Value (OMV) of residential property for owner occupation.

Despite the widespread use of the method in the UK, criticism has been levelled that DCC, as currently applied in practice, is imprecise and ambiguous (Wiltshaw, 1991) and weak (Jenkins, 1992). This concern followed the collapse of the housing market in the late 1980's but has lingered well into the 90's, leading one commentator to suggest that a "chartered surveyor is no more than an estate agent in a Sunday suit" (Anon, 1996a). More recent research (Gronow et al, 1996) has considered that both valuation practice and the structure of the market for residential valuation advice are at fault.

This paper presents a condensed critique of problems noted in practice (section 2). It suggests ways in which modifications can be made to practice to improve the accuracy and quality of advice supplied to clients (section 3) in particular through technology transfer (section 4). Finally it reflects on comments from psychologists on problems associated with professional decision making (section 5).

2. SOME PROBLEMS WITH CURRENT PRACTICE

The Red Book (RICS, 1995) outlines the role of the residential mortgage valuer as being to provide the lender (who is in most cases the client), with an estimate of OMV for the subject property. The process used in determining this value, DCC, is documented elsewhere (see for example Mackmin, 1994; Almond et al, 1997a). Given the widespread use of the approach the space devoted in texts to DCC is small. This may be a contributory factor in the poor application of this method in practice, which stems too from the minimal attention paid historically to residential valuation as part of surveying degrees.

A key issue in the valuation process is data. Given that information on sales is not readily available from the Land Registry or Inland Revenue in England and Wales, valuers must rely on their own valuations, those of colleagues, or potentially less reliable information supplied by third parties. This contributes to the situation in which the valuer has a limited number of comparable sales from which to draw an opinion of value.

Information on sale prices is critical, but weaknesses would still exist unless specific attributes relating to that transaction, which formed that opinion of value, were also available. Overall, there is a lack of consensus amongst practitioners as to which variables impact on value, with current practice placing an emphasis on supply-side (property related) attributes, a fact which is only too evident from reviewing mortgage valuation forms from lenders. This is the case, despite the fact that values in the market are derived through the interaction of supply and demand.

Further problems exist with how information is gathered, stored, accessed and applied in the appraisal process. While there is little doubt that leading residential surveying firms have been getting to grips with database technology, a survey of lending institutions, recently undertaken by the authors, confirms that technology take up is uneven. (43 responses were received from 81 questionnaires sent out, representing a 53% response rate). Comparables databases are common though not universal (and surprisingly absent in some of the bigger lenders). IT is limited, particularly for use within the inspection and appraisal itself. Few use or plan the use of computers in the field. Fewer yet use any form of statistical analysis and none of the respondents indicated that they used neural networks (though one smaller institution is planning their use).

Given the under use of databases, there is a lack of rigour in the selection process. Shortcuts appear to be made, with experienced valuers drawing on comparables from memory. It has also been shown from research in the US (Wolverton and Diaz, 1996) and the UK (Gronow et al, ibid.) that revealing to valuers the tentative sale price, as agreed between the buyer and seller, introduces a biasing effect in the selection of comparables and the resulting valuation.

Another danger is that valuers will be influenced by external pressures from clients. Tentative results from a recent US study (Worzala et al, 1996) suggest that residential appraisers were not likely to be so influenced. Nevertheless, comments received from practitioners by correspondence and at formal meetings suggest that client pressure may be a feature of the UK market place. Indeed, the valuer's own attitude was revealed in the case of *Gibbs and Another v Arnold Son and Hockley* (1989) 45 EG 156. Here, the defendant valuer stated that "valuation is an inexact science ..." (no difficulty in concurring with this statement) "... if the parties' agreed price is about right, it is irresponsible for the valuer to protect his own back by valuing at a slightly lower figure, because this means the sale may well go off, and the valuer's fee, which the buyer has had to pay, will effectively have been wasted". In this instance there is client pressure and the valuer is willing to acquiesce.

The key questions are, "how widespread is this pressure?" and "what defines "about right" in relation to the agreed price?" If Anon (1997) is reliable it appears customary in the UK for valuers to produce valuations at this tentative price, so as not to jeopardise the transaction, where the suggested value is to within 10-15% of the tentative sale price. Such practice is clearly not in the best interests of the purchaser, nor of the economy, and goes against suggested practice in the Red Book (RICS, ibid.), which states that the valuer's role is not to recommend the amount or percentage of the advance.

If comparable evidence is being selected to prove a value, the adjustment grid, as suggested by the texts, would be arbitrary and an inconvenience. This may account for the absence of formal adjustment processes by residential valuers in the UK. The cumulative impact of these effects could be aggravated further when valuers are placed under increased pressure from, amongst other factors, a greater workload. Claxton (1997) suggests that professionals, placed in such circumstances, are more likely to fall back on initial intuitive judgements, even when information received after this judgement appears to disconfirm the initial figure. Schön (1995) has observed that that professionals will often reach an initial decision, which although not correct, is felt by the practitioner to be "sufficiently correct" in the context of the situation, and no further action is taken.

Following the collapse of the UK housing market, a number of cases for alleged negligence arose from the time when the market was "on the turn". This situation has highlighted the need for indicators that forecast market changes. Current information, such as house price indices and data on transaction levels are "backwards" looking, and therefore only note a market change after it has occurred. Valuers often fail to react until well after cyclical break points yet they have not developed leading indicators to help minimise the potential damage.

3. SUGGESTED IMPROVEMENTS

In the drive for improving current practice a number of practical solutions exist, namely:

Improve the current "manual" process;
Consider alternative "intelligent" approaches;
Apply better practice with the wider use of "intelligent" systems.

In the context of this problem, consideration is now given to a number of areas where improvements can be made in the appraisal process, including revisions to current practice but also ways in which solutions exist through adopting simple IT approaches. Consideration

of "intelligent" approaches is provided in the following section.

Improvements to practice must be driven by the need for better data, in particular access to sufficient information on previous sales. The widespread use of DCC is on the basis that it provides the best opinion of value given the availability of data compared to other methods. Even where access is available within larger organisations such as Building Societies, problems exist with regard to the depth and reliability, with the under use of databases as noted earlier a key issue.

The main problem, particularly for valuers in England and Wales, is that information on sale prices is not currently available from the Land Registry, nor that of information relating to the transaction itself. This contrasts with practice in the US, where appraisers have access to Multiple Listing Services, databanks of sales information run by commercial organisations. In this respect it is suggested that a nationally maintained database be made available to all valuers, containing details on sale prices and information relating to the transaction itself. The Land Registry or the Inland Revenue would be a potential owner of the database, with a clear specification made by Government, consumer and professional bodies. The potential problems noted with commercially run services in the US (Crockham, 1995) should be avoided.

With regard to the information, the level of data that needs to be recorded is a subject on which practitioners do not readily agree. Current practice focuses on data relating to the property itself, which provides more information than is required in the context of the appraisal. What is needed initially is a classification of attributes which impact on value and the authors offer the classification in Figure 1 as a starting point for a discussion of this critical issue.

Beyond outlining the data important to the valuation, consideration must also be given to the way in which it is recorded, accessed and manipulated within the appraisal. The wider use of database technology to record and access data is required and standards need to be agreed in order to facilitate dissemination and analysis. The use of BS 7666 to standardise address data is a minimum. In Figure 2 we suggest the top-level structure of a data model for the inspection phase of the valuation process again in the interest of promoting a discussion.

The use of databases in the selection process is invaluable. Comparables can be selected using simple or complex pre-determined criteria. For properties with more unusual features a search can be made on a specific word to find any other properties with similar features; such a process in a manual system is almost impossible unless the valuer is able to recall such an inspection clearly from memory (limiting the search to personal experience). Of course, in selecting comparables automatically, selection bias is suppressed.

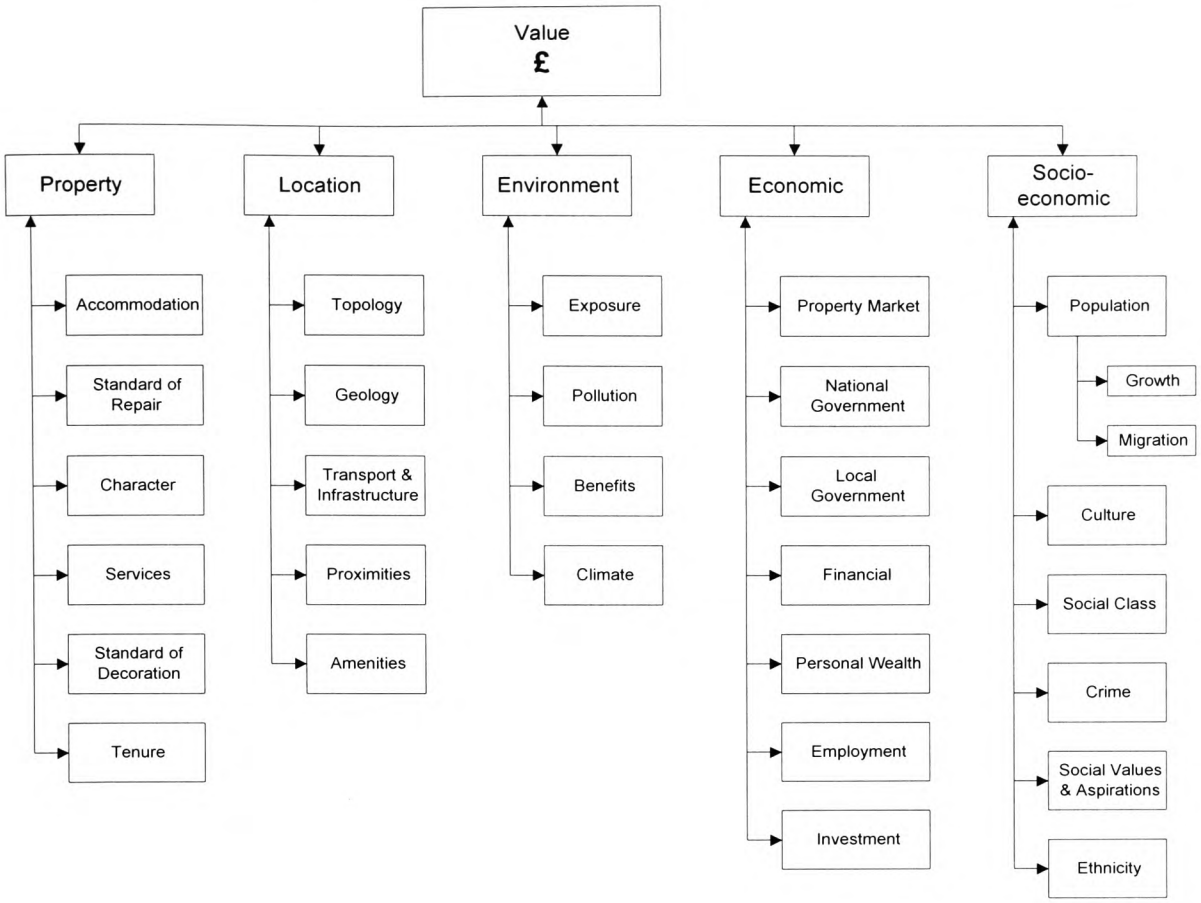
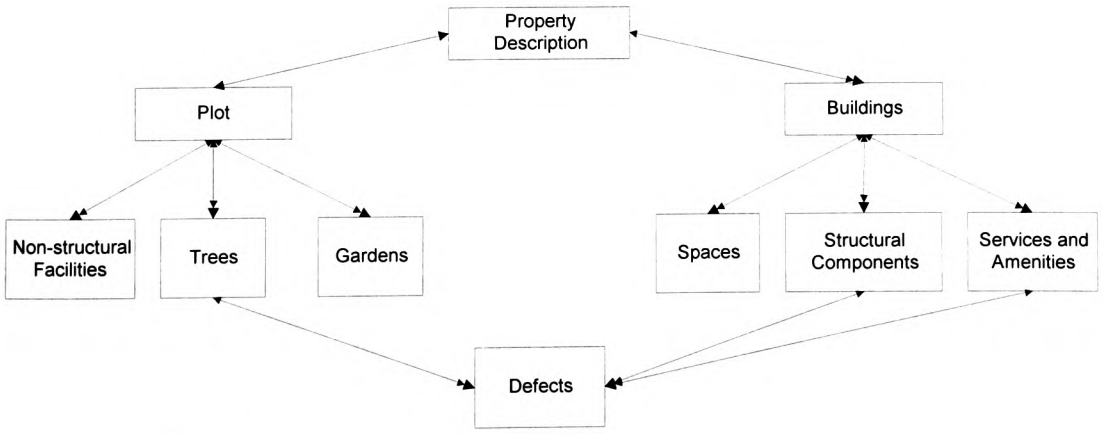


Figure 1. ***A Tentative Classification of Attributes Which Impact on Value.***

Guidance in the Red Book states that "cautious weight" should be given to the use of information that has yet to be realised in the market. On this basis it is suggested that the tentative price should not be provided to valuers (Almond et al, 1997b).



A double arrow denotes a one-to-many relationship.

Figure 2. ***A Suggested Data Model for a Property Inspection.***

Having accessed the necessary information, improvements are required to the way in which

this information is analysed and used in forming an opinion of value. The current process relies on the subjective opinions of the individual valuer. This is contrary both to theory, which suggests the use of an adjustment grid, and practice elsewhere. The authors are persuaded for the introduction of such practice in the UK, which provides clients, not only with details on which comparables were used, but also justifications for any adjustments made for differences in the comparable against the subject property. The wider use of demand side data should also be encompassed.

Given that a mortgage is more at risk within the early stages of the loan, where prices are more vulnerable to exogenous shocks in the market, it would appear logical to consider the future sustainability of prices within the market place. In this respect we concur with professionals who argue for the use of Estimated Realisation Price rather than OMV for mortgage appraisal purposes, though we go further and suggest that the real requirement is the application of predictive techniques first developed for commercial property (Connellan and James 1996).

In view of the problems associated with traditional practice, as well as changes to that practice, consideration need also be given to the development and use of alternative "intelligent systems", to which we now turn.

4. INTELLIGENT SYSTEMS IN THE VALUATION PROCESS

At the most basic level, databases provide a useful tool in the appraisal process, given that the appraisal is driven by the need for data. Despite the availability, and ease with which databases can be set up and valuers trained in their use, a combination of both the recession and cultural reasons long delayed their use (Almond et al, 1996). Indeed it is notable that where systems have been implemented, it is the administrative and not the appraisal processes that have come first signifying that business management not professional effectiveness was the driving force.

This is a far cry from the situation postulated a decade ago by Gronow and Scott (1987). The technical means already existed in the valuation of residential property for technicians to perform inspections, databases to access comparables, expert systems to underpin the valuation itself, whilst reports could be automatically produced using standardised forms.

Despite the poor up-take of IT by the profession a number of practical solutions do exist in prototype, to show the potential that exists with new technology. Outlined within this paper are systems for the collection of data on-site, the use of database technology in the analysis and formulation of value, and the use of Artificial Neural Networks (ANN's) to form an additional opinion of value.

4.1. ON-SITE DATA CAPTURE

The consideration of systems to capture data on-site is not new. The use of such systems has been constrained by practicalities such as cost, and the size, weight and weather-resistance of hardware for valuers to use within the field. To date most financial institutions have concerned themselves in reducing the time spent in transmitting data electronically. For example the Britannia uses e-mail allowing valuers to work at home, but also has a system whereby inspections can be dictated over the phone to their Head Office from on-site (Anon, 1996b; Eade, 1996). The authors survey, considering the use of IT, did note some institutions using IT in the field, but it is very limited in terms of scope.

The aim of such systems may be to reduce the time spent on the inspection itself, though consistent, standardised reporting in electronic format may rather facilitate report quality. If

systems are to be used to capture data on-site they must be small, light and portable, with a screen large enough to view and capture data, with easy to use controls.

The advent of Windows CE™ which runs on the latest HPC's (handheld personal computers) provides a useful platform with which to capture data. The hardware is light, comfortable and small (measuring for example some 160mm x 90mm). The software is the familiar Windows style front-end (which is widely known and user friendly), which also means data can be easily transferred onto a standard PC. Beyond this, it is also some ten times cheaper to acquire a HPC than a less portable "laptop".

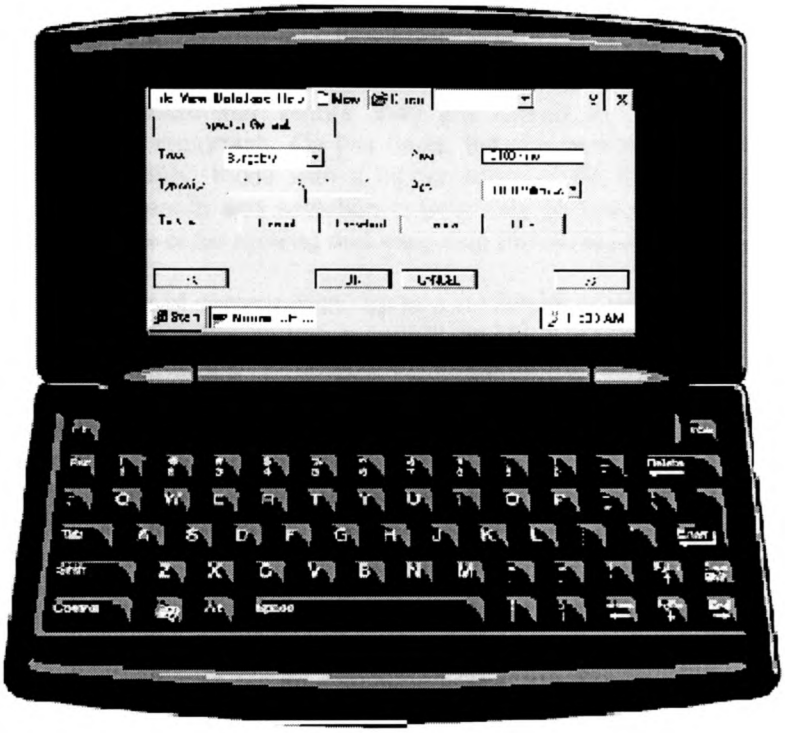


Figure 3. Example of a HPC.

At the University of Glamorgan the research and development team have produced a prototype system, programmed using C++. This is not a do-it-yourself development environment: it requires professional programming skills. Although it can be expected that productivity tools will facilitate the development of bespoke software for the HPC, such products that have emerged to date (e.g. FormLogic) also require knowledge of C++. An example screen from the prototype is shown in Figure 3, from which it can be seen that data can be captured in a number of ways; the use of a conventional keyboard, or a "built-in pen" to click on boxes or to select a pre-defined keyword from pull-down menu's. A slider bar is also provided which provides for fuzzy decision making within the appraisal process.

Such systems enable sufficient data to be captured from an inspection for the purpose of a simple mortgage valuation, and can also act as an *aide memoir* to the valuer, providing a warning if any important aspect of the inspection has been missed. Ideally data will be transferred to a database on a desk-top PC from which analysis can be performed before the compilation of a report. However, the provision of Pocket Word permits a report to be created using standardised paragraphs within the system, and sent via mobile phone to the office.

4.2. EXPERT DATABASE SYSTEMS

Expert database systems contain or interface to the data that is required for the decision-making process. Additionally they mimic the valuer in respect of routine aspects of the appraisal process. There is a considerable literature on the development and use of such systems (Gronow and Scott ibid., Scott 1988, Jenkins ibid.).

A prototype system, developed as part of a Realising Our Potential Award from the ESRC, automatically selects comparables, through which an adjustment grid style display allows the user to make manual adjustments for noted differences (see Figure 4).

The automatic selection of comparables is on the basis of a simple matching process in which attributes of the subject property are matched against all other properties in the system; those properties with the greatest number of matches are selected. Of course, given that certain attributes are more important than others, they are ranked and weighted and the search ripples out from the more important. On this basis, the comparables returned by the system are in order of comparability, those with a higher score come first and so on. Given that valuers use a range of search and selection criteria, any commercial system would require that the valuer is able to re-order ranking and weighting (not a feature of the prototype).

Having returned a number of comparables, up to a maximum of ten, the system provides the user with the ability to refine the search manually on the basis of a number of criteria, i.e. Property type; Unit type; Number of bedrooms; Value range; Valuation date; and Location (street postal area, district, town, city or county). A similar but more exhaustive process can be used from which to select a candidate as a "matching pair". From the comparables returned, the user can browse across and down the properties, and view their respective details (attributes) to consider the level of comparability. The user will then select the most appropriate comparables (up to five) which will be used for further analysis to form an opinion of value.

The adjustment process (see Figure 4) allows the user to scroll down and across the comparables selected and note, by clicking in the appropriate box, a value adjustment for the noted difference, for example in Figure 4, a hypothetical adjustment of £500 is made for the difference in floor area. Once an adjustment is made, the value estimate is automatically adjusted. Given the differences of time between the valuation date of the comparable and that of the subject property, a house price index is applied to smooth the value estimate automatically, a process which valuers rarely make explicit, yet important in terms of market changes. Of course, such adjustments are blunt instruments and a commercial system would require that control be given to the valuer to overwrite defaults.

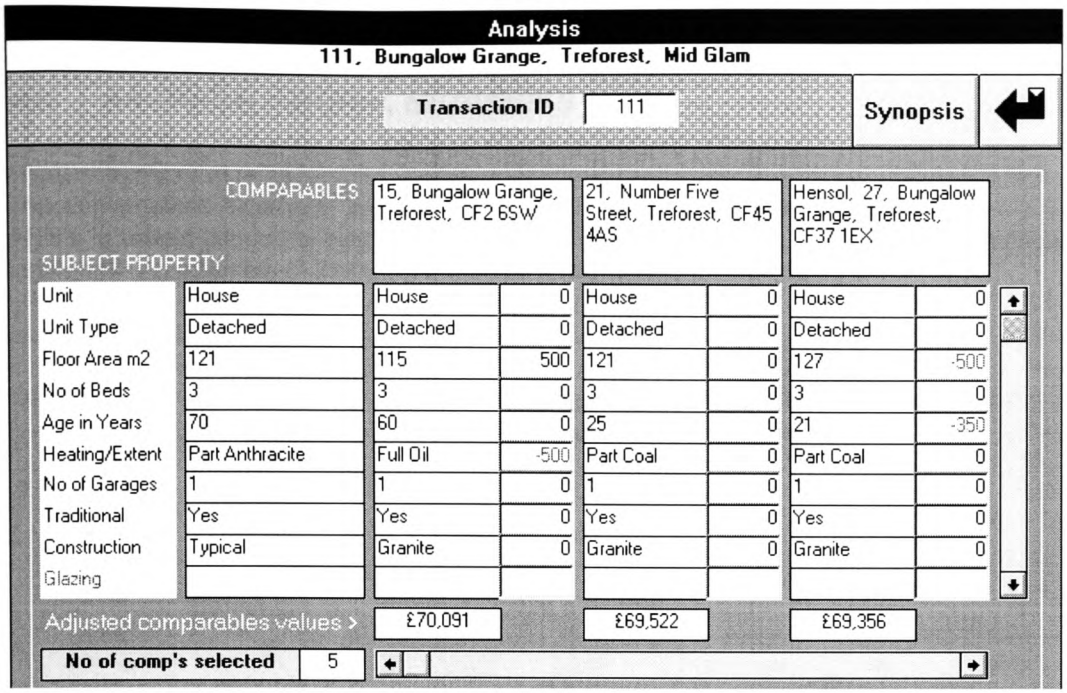


Figure 4. Comparables Adjustment.

On completion of the adjustment process, the system provides an analysis of the data (see Figure 5). The information displayed includes the value of the closest match, together with estimates of the individual comparables, which form a range within which the value of the subject property is likely to lie. Additional information such as value per square metre and house price index are also provided. There is also additional space to allow the incorporation of a neural network value at some further stage.

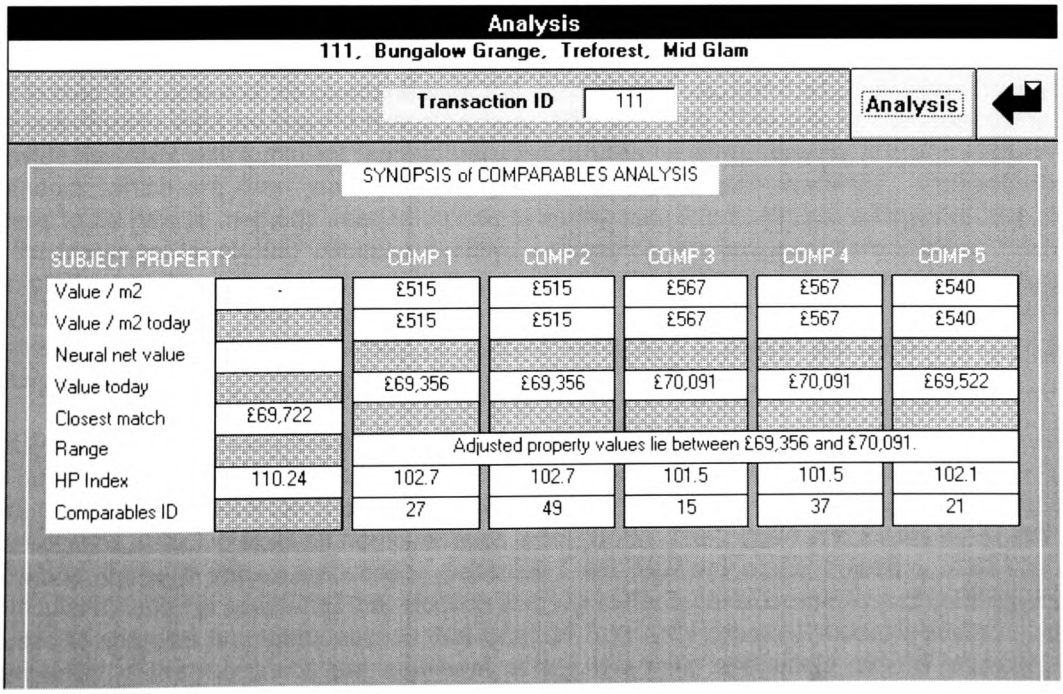


Figure 5. Comparables Analysis.

4.3. ARTIFICIAL NEURAL NETWORKS

The conventional approach used to build an appraisal model for residential property using ANN's is based around a single Multi-Layered Perceptron (MLP) using a Back-Propagation algorithm. An example of this simple type of network is shown in Figure 6.

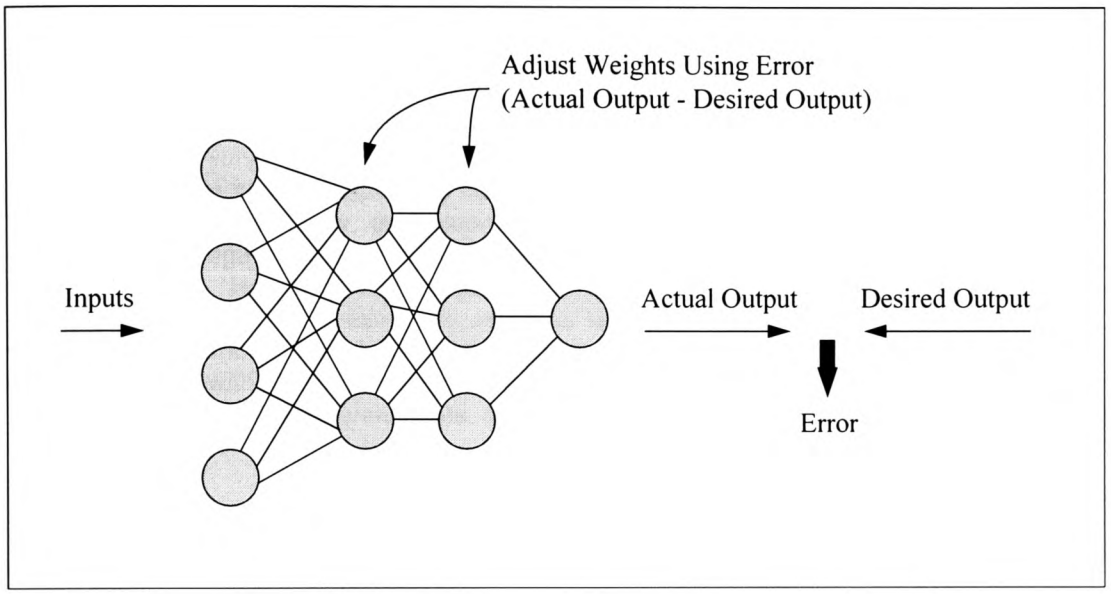


Figure 6. A Simple Multi-Layered Perceptron Supervised Learning Architecture.

Building such an appraisal model requires a number of tasks to be performed. The first and the most important of these is pre-processing of the data. This operation needs to be performed by someone who understands the domain and has sufficient knowledge to select the appropriate data for the modelling process (Bigus, 1996). The aim of this exercise is to provide the ANN with sufficient representative examples, in an acceptable format (typically all numeric), which are most susceptible to accurate and unbiased modelling. Consideration needs to be paid to methods used to recode symbolic data into numerical surrogates, but also to the treatment of missing values, subjective information and untypical transactions. Having pre-processed the data, it is important that the data set is divided into at least three subsets (Bigus, ibid.), one to train the network (training set), one to test the network during training (validation set) and at least one hold out sample to be used to assess the performance of the network (test set).

Published research using this technique, normally using only property specific attributes such as number of bedrooms, floor area, and house type, typically show a percentage difference of less than 10% between the predicted values, and those returned by the valuer, for mortgage transactions selected from an homogeneous area (Borst, 1991; 1994 and Evans et al, 1992). However, property values are location dependent, and thus any model trained on data from a particular location is specific to that location and will suffer a considerable loss of performance if used to appraise properties outside that location. The ANN appraisal model is limited by the complexity of the function it can represent at any one time, and would require retraining to achieve the new objective.

Hence, it follows that a computer appraisal system based on this ANN technique would require a large number of dedicated appraisal models in order to simulate the appraisal

process over an heterogeneous area. It is the automation of this process and the discovery of the physical boundaries, whereupon one ANN model is replaced by another, which is of most interest. Potential sources for this information may be contained within the numerous commercial geo-demographic indication systems, and also the UK Census data, on which the indicators are based.

4.3.1. Description of Census Data

The 1991 UK Census provides researchers and the Government with the "most authoritative social accounting of people and housing in Britain" (Dale and Marsh, 1993). Comparable statistics are generated for very fine geographical areas, the smallest of which is an Enumeration District (ED) in England and Wales, and an Output Area (OA) in Scotland.

Census data have been included in a number of projects concerned with residential property appraisal. Most significantly, professionally constructed geo-demographic indicators based on Census aggregates, are included in the house price index issued by the Nationwide Building Society. However, the characteristics presented by the indicators are not always representative and although extremely useful for targeting and advertising customised goods, they are more general than the underlying property markets contained within. It is for this reason, coupled with the fact that commercial geo-demographic indicators are not as readily available to academics as Census data, that this study focuses on raw data collected from the 1991 UK Census.

4.3.2. Analysis

Two methods were considered for adding Census data into the appraisal model. The first simply required the selected Census aggregates to be added to the attributes used to train the single ANN model. This coupling was achieved using a postcode to ED cross-reference file. The results of this analysis are shown in Table 1.

Table 1. Results Obtained for ED level Analysis.

Data Sample	Mean Absolute % Error	Improvement
Sample of Cardiff Data set	20 %	-
Cardiff Data set and ED level Census statistics.	13 %	7%

Clearly, a marked improvement was achieved when the Census data was included. Moreover, a detailed study (Gronow et al, ibid.) has revealed that the location size represented by the Census data has a direct influence on the effectiveness of the model. It also notes that regional characteristics are made up of many factors, including obvious influences such as housing stock, tenure and employment, but also containing less obvious factors such as occupation and number of cars per household.
For a single ANN model of this type to successfully simulate the appraisal process within an heterogeneous area, it must be assumed that there is one valuation function shared across all homogeneous sub-areas, which can be adjusted along a sliding scale according to the economic characteristics of each area. However, the very nature of the heterogeneous housing market does not consistently support this assumption. Faced with this situation, Adair et al (1996) hypothesised that sub-markets could be identified by stratifying the housing market into increasingly homogeneous data sets. Extending this argument, it is logical to expect that independent modelling of each sub-market would permit the ANN model to be more specific and therefore more accurate in appraising properties selected from within a sub-market. The stratification of the housing market using Census data was the aim in the next

section of research.

4.3.3. Stratifying the Housing Market

The purpose of stratifying a data set is to decompose it into more manageable sub-sets. This general idea has been used to compute theoretical distribution models using statistical methods for more than 50 years. In this analysis, stratification is achieved using a technique called the Kohonen Self-Organising Map (Kohonen, 1984), of which Figure 7 is an example.

Each node on the feature map contains a vector of length 'j', where 'j' is equal to the number of input dimensions or features. Before training, the network is in an initialised state (i.e. the directions of the vectors in each node are random). Training involves passing an input vector into the network through the input nodes. Each node on the feature map is then compared with the input vector, and the closest node is then changed to be more like the input vector. Neighbouring nodes also become more like the input vector. Iterating this process achieves spatial clustering of similar input vectors.

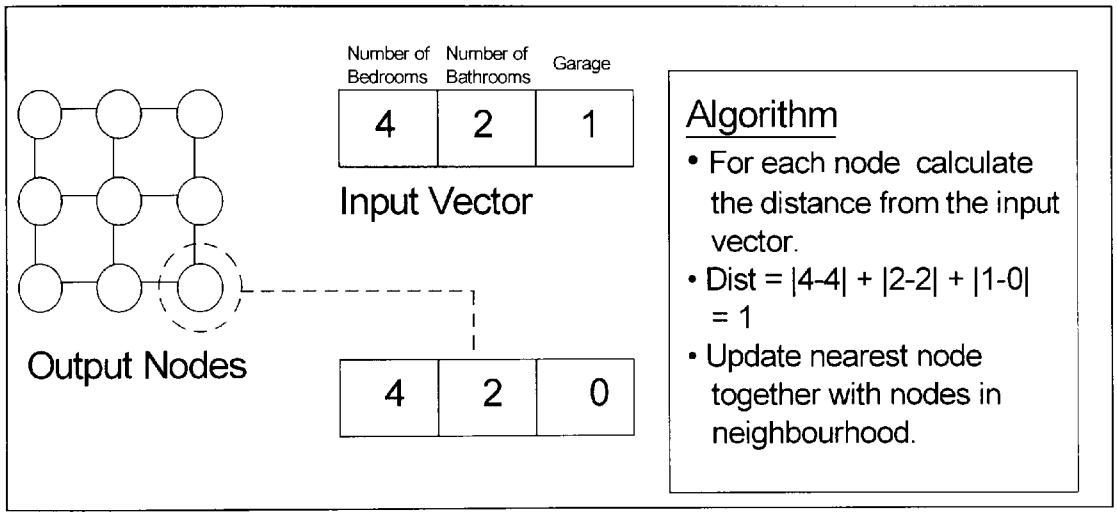


Figure 7. A Kohonen Self Organising Network.

The methodology involves training the Kohonen network on locational data and using the observed 'clusters' as training sets for MLP networks of the type shown in Figure 6. The Kohonen network is trained on Census aggregates, extracted at the ED level. The data from each significant grouping (formed by the Kohonen network) are cross-referenced using an ED to postcode file with mortgage transaction data to form sub-populations. A separate MLP network is trained for each sub-population. Figure 8 describes pictorially a framework for integrating these two techniques.

The Kohonen Self Organising Map has been used in a number of studies for market segmentation (Bigus, ibid.; Openshaw and Wymer, 1995). These implementations focus exclusively on the Census data during clustering and aim to provide general geo-demographic indicators. However, a study (Lewis et al, forthcoming) highlights the importance of selecting clusters that match the functionality of the domain data. The fact that similar characteristics are shared by a sub-population does not necessarily mean that the sub-population represents an homogeneous area with respect to property prices. In order to overcome this problem, sub-populations or clusters, were selected using an algorithm that estimates the ability of a neural network to learn the underlying functions present in the described region (Lewis and Ware, 1997). Clusters with large error estimates were not used as training sets for the suite

of MLP networks. In addition to these sub-models, a single model was created using all of the training data intended to predict any properties not covered by the sub-models.

In order to test the effectiveness of the method for determining useful residential property sub-markets, two testing procedures were developed. The first concerned the ability of the sub-models to outperform a single model trained on all of the domain data. Sub-models were constructed based on individual geo-demographic factors, for example *housing stock in region* and *employment statistics*. Test properties were passed through the single neural network model trained on all of the available data and also through the appropriate sub-models. The property value estimates made using this technique were between 1% and 14% closer to the values returned by the valuer than were the corresponding estimates made by the single ANN model.

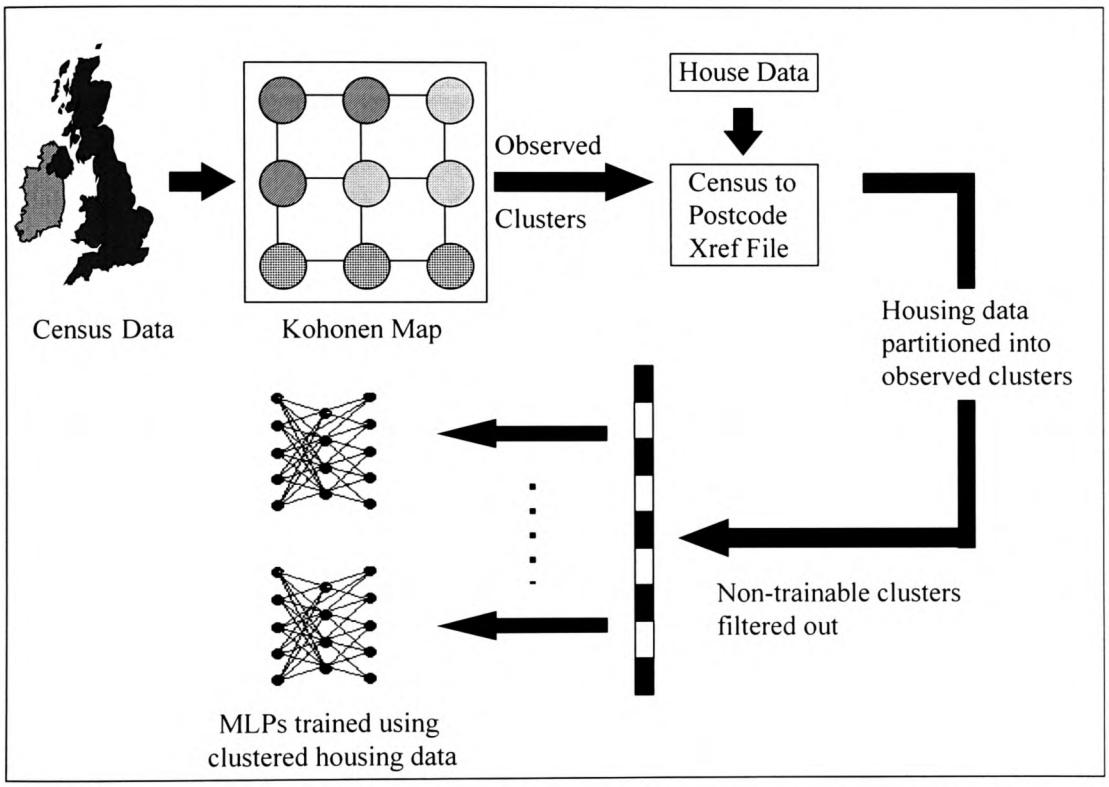


Figure 8. ***A Framework for Integrating the Kohonen Map with a Suite of MLPs.***

The aim of the second testing criteria was to determine whether a sub-model trained using data selected from one geographical area could effectively be used to predict the values of properties from a different geographical area, a feat that the single model trained on homogeneous data cannot currently achieve. ANN models were created using property data selected within a sub-market from one county in South Wales (County_A) and were used to predict the values of residential properties in a similar sub-market in a neighbouring county (County_B). The predictions made by the County_A models were compared with predictions made by a single neural network model trained on all of the County_B properties. The results (see Table 2) show that the County_A models at least match, and in some cases outperform, the predictions made by the single neural network model trained on County_B properties.

Clearly, the ability of these models to predict the values of residential properties outside the geographical area from which the training data was selected has been demonstrated. However, using the Kohonen Self Organising map to stratify the data into homogeneous training sets is not always optimal as success is highly dependent on the choice of variables

presented to the network. To enhance this process it is proposed that a more flexible approach to variable selection is adopted where variables are introduced based on their incremental gain using a multi-dimensional tree structure. Having achieved trainable clusters at nodes on the tree, descriptive rules could be extracted by tracing down appropriate branches. These rules could then be compared with valuation knowledge as well as rules induced using other rule induction techniques such as ID3 (Quinlan, 1986). The aim of this work, to be completed in 1998 as part of a PhD is to develop a framework for combining artificial intelligence, traditional statistical methods and appraisal knowledge to form an intelligent appraisal model.

Table 2. ***Sample of Results for the Two-County Analysis.***

Data Sample	Mean Absolute % Error	Improvement
County_B Whole Data set	18%	-
County_A Model_A	18%	0%
County_A Model_B	17%	1%
County_A Model_C	15%	3%
County_A Model_D	15%	3%
County_A Model_E	18%	0%

4.4 INTELLIGENT HYBRID SYSTEMS

Intelligent techniques such as Neural Networks, Expert Systems and Expert Database Systems have shown the potential to effectively model distinct parts of the residential property appraisal process. However, in complex domains, the weaknesses of each of the single intelligent techniques are exposed. Neural networks are good at representing underlying patterns within data, but have weak explanatory powers; expert systems and expert databases are good at representing cognitive decision processes but only respond appropriately under very narrow domains (Holland, 1986).

To overcome these component limitations, a new philosophy is emerging within artificial intelligence which focuses on the development of Hybrid Intelligent Systems (HIS), that are able to compensate for the weaknesses of the component modules through a co-operative approach. Goonatilake and Khebbal (1995) suggests that HIS exist in three primary forms:

Technique enhancement - where part of an intelligent technique that is weak is replaced by a corresponding and more effective part from another technique. For example, replacing the strict pattern matching of an expert system with a neural network (Schreinemakers and Touretzky, 1990).

Intercommunicating HIS - natural splits in the domain are dealt with by the most appropriate technique. For example, precise and well defined knowledge are represented by an expert system and ill-defined fuzzy data are modelled by a neural network (Corkill, 1991).

Polymorphic HIS - techniques are used to emulate the processes normally performed by a different technique. For example, symbolic reasoning within conventional neural networks (Hughes, 1992).

From this emerging field it may be possible to construct a framework that facilitates the complementary development of an intelligent appraisal system, a tentative model of how this might be applied to real estate is provided in Figure 9.

The development of such a hybrid system has two benefits. Firstly a better understanding of valuers decision-making processes can be achieved, e.g. the ability to elicit rules from a neural network (see Andrews and Diedrich, 1996) can provide an understanding of the pattern recognition process and an explanation of a property's value. Secondly, a better understanding of professional knowledge, such as that currently being developed at the University of Glamorgan (see Almond, 1997), will assist in developing more appropriate models, considering data important in the appraisal process, rather than data chosen on *a priori conjecture*.

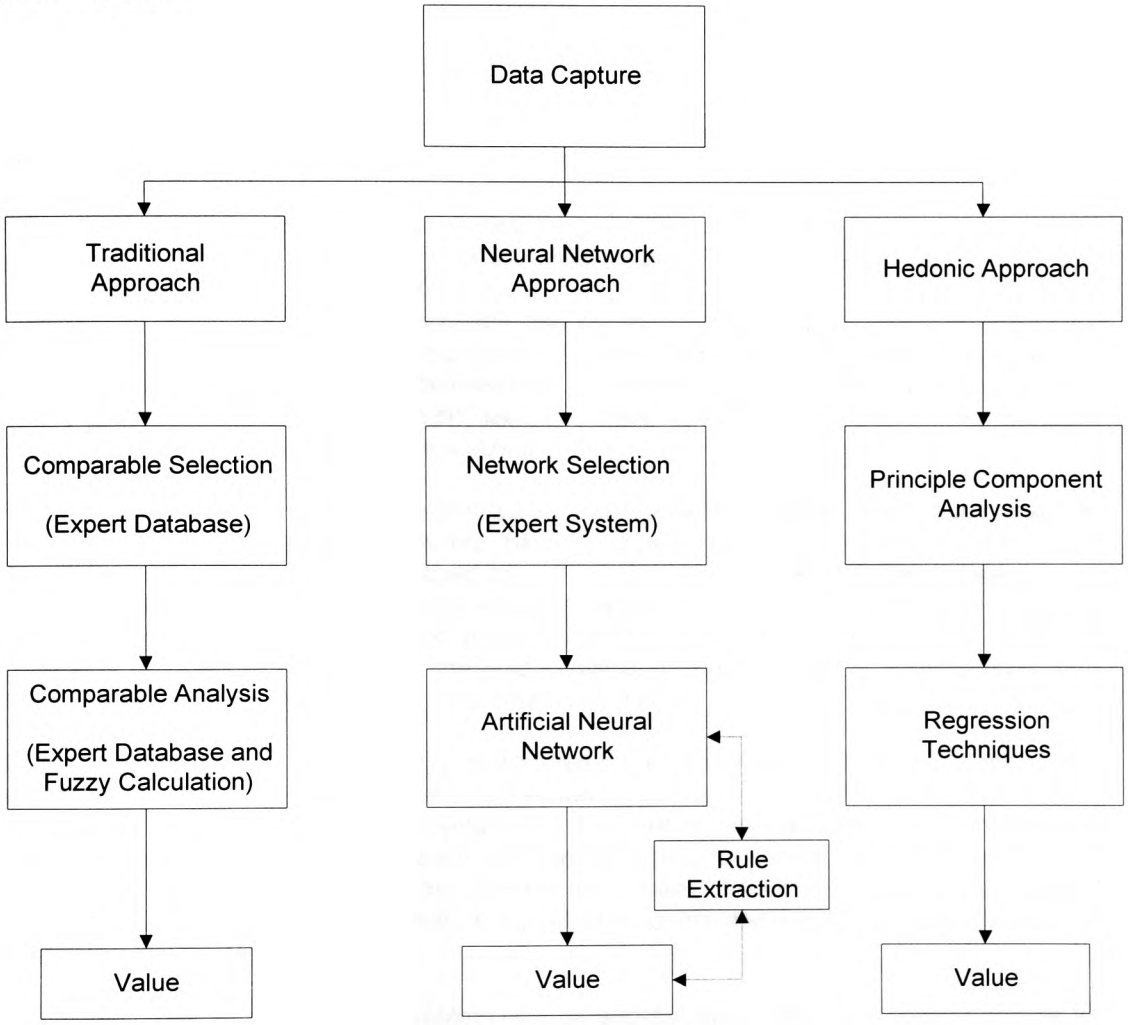


Figure 9. Tentative Structure of an Intelligent Hybrid System.

The hybrid intelligent system approach adds value to a whole range of applications, and has been considered by both academics and valuers of commercial software. Amongst the most promising are Blackboard Systems, see for example Corkill (ibid.), a technique where appropriate knowledge-based and statistical paradigms operate on discrete parts of a problem and share derived knowledge. Another alternative is a Hybrid Expert Network (Caudill, 1990) in which neural networks are embedded in traditional expert systems in order to action non-distinct "if" clauses.

Commercial HIS development software is beginning to become available that will facilitate the integration of cognitive knowledge with machine learning paradigms. Progress in this field of research could result in an intelligent appraisal system that combines both the knowledge and

experience of professional valuers with the data analysis prowess of current intelligent techniques.

5. PROFESSIONAL VALUATION KNOWLEDGE

The problems of modelling heterogeneous markets highlight the complexity of the residential appraisal process and the consequent need for a better understanding of the professional valuer's knowledge if the type of systems described in Section 4.4 are to be developed and refined. Perhaps the application of theoretical principles and techniques from behavioural psychology holds the key. Certainly, problems of professional knowledge are not confined purely to the real estate profession. Schön (ibid.) notes that in the general day-to-day professional world, practitioners continually rely on tacit recognitions, judgements and skilful performances. In this respect he highlights a crisis of confidence in professional practice, which is described as a mismatch between the professional body of knowledge and the ever changing professional world, in which decisions have to be made in the face of complex information, uncertainty and unique situations.

The problem is rooted in an allegedly false counterposition between theory and practice. Universities and professional bodies are meant to provide the theoretical basis upon which practitioners are then supposed to build an edifice of professional knowledge. However, this division between the theoretical and the empirical, which originates in pure sciences, does not travel well into social sciences and professional domains. Eraut (1994) shows how the division is reflected in surveying by the dual qualification system used in training surveyors; that of a degree followed by a period of professional training.

In residential valuation, the artificiality of the division is clearly exposed. There is no or little theory of direct and specific relevance to residential valuation in University courses. This is not to say that competence in surveying is not a matter for high academic and professional standards. Sound residential appraisal practice is socially beneficial. Students need to be equipped with a wide range of skills drawn from many disciplines. But, in the case of residential appraisal per se, very little is taught to students and there are very few texts of direct relevance.

Within the traditional division, the problem may be addressed by creating a specific degree for residential appraisal that encompasses techniques drawn from urban/ housing economics, geography and statistics. But Schön is more radical.

Schön's response to the situation is that of a reflective practice. He suggested that practitioners should "openly" reflect on the practice situation, i.e. enter into a dialogue with the practice situation, by reflecting both in, and on, action; surfacing ideas when stimulated by surprise and re-framing the situations according to circumstances, to consider future actions. Such reflection could be on technical aspects of practice, on the "lifeworld" in which the practitioner practices, the social, political or economic aspects of practice and the knowledge used itself (James, 1997).

Claxton, agreeing, (ibid.) states that "good learning requires the ability to be reflective; to take a strategic, as well as tactical, perspective on ones learning and knowing; to be aware of how things are going, and of what alternative approaches there might be". In this respect, the adoption by the profession of Continual Professional Development (CPD), promoting the idea of lifetime learning, was a positive step. However, the notion of "keeping abreast" that

permeates CPD ensures an emphasis on the legal-technical aspects of professional work. Yet, CPD could be enhanced to incorporate reflective practice, with both individuals and businesses engaging in what can be considered "corporate reflection", wherein practitioners can regularly reflect and engage in self and peer group knowledge evaluation.

Eraut (1995), whilst welcoming Schön's contribution, is also critical of his ideas, noting amongst other things, the problems of time and opportunity for reflection. Claxton (ibid.) too is aware that professionals are often faced with having to make judgements without a moment's thought. So much time is spent processing information, meeting deadlines, and solving problems that the practitioner is often left with no time to think.

According to Claxton (ibid.), the problem is not helped by new computers which process greater amounts of information in shorter spaces of time; this, it is suggested, stimulates the professional to think faster yet. In this respect Claxton suggests that the professional should take time to "mull over things" i.e. practitioners should "pull off the Information Super-Highway into the Information Super Lay-By; to stop chasing after more data and better solutions and rest for a while". This is to confuse the results of computer development with the business aims and objectives that lie behind them. Speeding the processing of data and knowledge per se creates greater time for reflection. But if the purpose of software development is simply to increase productivity opportunities will be lost for enhancing practice, for discovering the nature of professional reflection and for better understanding a valuer's knowledge and decision-making processes.

The prerequisite of improved practice is that professionals be adaptive to change rather than conforming to the status quo. It is the belief of Challinor (1997) that the surveying profession's poor response to technological and economic change has led to accountants taking work of surveyors.

One aspect of professional knowledge that is of immediate interest is that of local real estate valuation knowledge. Concern has been expressed (anon, 1993) that valuers travel "good" distances to little known locations to provide a valuation. This concern has bee recognised by the RICS. A prerequisite for any valuer performing a valuation, outlined in the Red Book (RICS, ibid.), is that the valuer should have "sufficient current local ... knowledge of the particular market and the skills and understanding necessary to undertake the valuation competently" (Practice Statement 5.1.a).

However no definition of the content of local knowledge is provided. In other literature, mention of local knowledge is limited. However it is interesting to note the degradation of performance of intelligent systems when they have been applied to secondary locations within the same region (Scott ibid.). And it is equally interesting to note that the winner of the IAAO 1996 Mass Appraisal Contest drew substantially on local knowledge.

It is assumed that intelligent systems will incorporate local knowledge and that their most efficient deployment will be by skilled appraisal personnel with local knowledge. The prerequisite is that there is an understanding of the content of local knowledge and how it is processed.

Research is currently underway to investigate in more detail the issue of local valuation knowledge by way of a knowledge elicitation exercise with valuers. It is already understood from our previous observations of practice that valuers may use market data as cues to the performance of the market. The aim of this work, to be completed next year as part of a PhD will be to consider if such information is used, and, if so, what type of data is noted, and what use it has in the valuation process (Almond, ibid.).

6. CONCLUSIONS

This paper has critically considered the current application of DCC in regard to residential mortgage valuations. The concern is that unless changes are made to practice the consequences of the late 1980's property crash may well recur. Given this situation a number of solutions to practice have been put forward including the use of new technologies. The vision is that such systems will at first aid the valuer in reducing the amount of time spent on the more mundane aspects of the valuation process, enabling the valuer to focus on the more important aspect, the appraisal itself. This would not only lead to greater accuracy, but also provide better information to clients. The degree of substitution for professional skills in future depends critically on an understanding of professional knowledge.

The "intelligent" approaches outlined are still in the development stage. Software has been written for valuers to capture data on-site and undertake routine analysis using expert database systems. The integration of neural networks into such systems is possible, though the preliminary development of an intermediate stage in which appropriate networks are intelligently selected is recommended

Empirical research shows that the addition of Census data at the ED level into an ANN appraisal model significantly increased its accuracy. However, with consideration to a priori knowledge relating to the varied interplay of demand and supply side variables across different geographical regions, it was concluded that the Census data could be more effectively employed as a method of segregating the heterogeneous property market into homogeneous sub-markets.

Given this, a technique was developed in which Census clusters were used to describe the content of a collection of training sets that were each modelled independently using an MLP network. The outcome of this suggests that:

A set of models, each dedicated to a certain narrow domain, can significantly outperform predictions made by a single more general model trained on all of the available training data. Models created from the stratification technique can be used to predict property values in other areas that have similar Census characteristics.

Based upon these results consideration has been made to the development of hybrid systems, where different technologies can be brought together into one system to compliment each other. However, the introduction of intelligent systems requires a deeper understanding of valuation knowledge and decision-making processes.

REFERENCES

Adair, AS, Berry, JN and McGreal, WS, 1996, Hedonic Modelling, Housing Submarkets and Residential Valuation, *Journal of Property Research*, Vol. 13, 67-83.

Almond, N, 1997, *The Development of an Holistic Methodology for the Valuation of Residential Property*, Progress Report for Transfer from MPhil to PhD, University of Glamorgan.

Almond, N, Gronow, S and Jenkins, D, 1996, Applying IT to Valuation, *The Valuer*, November/ December, pp 22-23.

Almond, N, Jenkins, D and Gronow, S, 1997a, *A Comparative Study of Residential Valuation Techniques in the UK*, Fourth European Real Estate Society Conference, Berlin, June 25-27.

Almond, N, Gronow, S and Jenkins, D, 1997b, Sound Valuation Evidence, *The Valuer*, May/ June, pp 24-25.

Andrews, R and Diedrich J, 1996, *Rules and Networks*. Proceedings of the Rule Extraction from Trained ANN Workshop, AISB96

Anon, 1993, Estate Agents Slam Low Valuations, *Estates Gazette Interactive - EG Archive*, 14 August.

Anon, 1996a, Surveyors Wallow in Doom and Gloom, *Daily Telegraph*, 22 October, p 27.

Anon, 1996b, Home Work, *Mortgage Finance Gazette*, June, p 52.

Anon, 1997, Scrutinise Those Surveyors' Valuations: Letters to the Editor, *Property Week*, 6 June, p 17.

Bigus JP, 1996, *Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support*, McGraw-Hill.

Borst, RA, 1991, Artificial Neural Networks: The Next Modelling / Calibration Technology for the Assessment community?, *Property Tax Journal*, Vol. 10, pp 69-94.

Borst, RA, 1994, *A Method for the Valuation of Residential Properties using Artificial Neural Networks in Conjunction with Geographical Information Systems*, IAAO Conference, Dublin.

Caudill, M, 1990, Using Neural Networks: Hybrid Expert Networks, *AI Expert*, Vol. 5 (11), pp 49-54.

Challinor, P, 1997, Losing Ground to the Outsiders, *Estates Gazette Interactive - EG Archive*, 23 August.

Claxton, G, 1997, *Hare Brain, Tortoise Mind: Why Intelligence Increases When You Think Less*, Fourth Estate.

Connellan, OP and James, H, 1996, *Estimated Realisation Price by Neural Networks*, RICS Cutting Edge.

Corkill, D, 1991, Blackboard Systems, *Artificial Intelligence Expert*, September Issue.

Crockham, J, 1995, Sales Comparison Approach: Revisited, *Appraisal Journal*, Vol. 63, pp 177-181.

Dale, A and Marsh, C, 1993, *The 1991 Census User's Guide*, HMSO Publications.

Eade, C, 1996, Home Help, *Property Week*, 28 March 1996, pp 38-39.

Eraut, M, 1994, *Developing Professional Knowledge and Competence*, Falmer Press.

Eraut, M, 1995, Schön Shock: a Case for Reframing Reflection-in-Action?, *Teachers and Teaching: Theory and Practice*, Vol. 1(1), pp 9-22.

Evans, A, James, H and Collins, A, 1992, Artificial Neural Networks: an Application to Residential

Valuation in the UK, *Journal of Property Valuation and Investment*, Vol. 11, pp 195-204.

Goonatilake, S and Khebbal, S, 1995, Intelligent Hybrid Systems: Issues Classifications and Future Directions, in *Intelligent Hybrid Systems*, eds. Goonatilake, S and Khebbal, S, John Wiley and Sons.

Gronow, S and Scott, I, 1987, Information Technology and Building Society Valuations, *The Valuer*, March, p 58.

Gronow, SA, Ware, JA, Jenkins, DH, Lewis, OM and Almond, NI, 1996, *A Comparative Study of Residential Valuation Techniques and the Development of a House Value Model and Estimation System*, ESRC End of Award Report No.RO222500045.

Holland, JH, 1986, Escaping Brittleness: The Possibilities of General-Purpose Learning Algorithms Applied to Parallel Rule-Based Systems, in *Machine Learning 2* , ed. Michalski R, Carbonell J, Morgan Kaufman.

Hughes, C, 1992, *Neural Networks and Expert Systems - A Partnership*, IEE Colloquium of Neural Networks and Knowledge Based Systems.

James, C, 1997, *How do you do? An Introduction to Professional Knowledge and its Development*, University of Glamorgan.

Jenkins, DH, 1992, *Expert Systems in the Land Strategy of Cardiff City Council* ,Unpublished MPhil Thesis, Polytechnic of Wales.

Kohonen, T, 1984, A Simple Paradigm for the Self-Organised Formation of Structured Feature Maps, in *Competition and Co-operation in Neural Networks*, eds. S. Amari, M. Arbib. Vol. 45. Berlin: Springer Verlag.

Lewis, OM and Ware, JA, 1997, A Novel Neural Network Technique for Modelling Data Containing Multiple Functions, in "Lecture Notes for Computer Science" series 1226, *Computational Intelligence: Theory and Applications*, Springer Verlag.

Lewis, OM, Ware, JA and Jenkins DH, forthcoming, A Novel Neural Network Technique for the Valuation of Residential Property, *Journal of Neural Computing and Applications*, Springer Verlag.

Mackmin, D, 1994, *The Valuation and Sale of Residential Property*, 2nd Edition, Routledge.

Openshaw, S and Wymer, C,1995, Classification and Regionalisation, in Openshaw, S (eds.) *Census Users' Handbook*, Longman, London.

Quinlan, JR, 1986, *Induction of Decision Trees*, Machine Learning, Vol. 1, pp 81-106.

Royal Institution of Chartered Surveyors, in association with the Incorporated Society of Valuers and Auctioneers, Institute of Revenues Rating and Valuation, 1995, *RICS Appraisal and Valuation Manual*, RICS Business Services Ltd.

Schön, DA, 1995, *The Reflective Practitioner: How Professionals Think in Action*, Arena.

Scott, IP, 1988, *A Knowledge Based Approach to the Computer-Assisted Mortgage Valuation of Residential Property*, Unpublished PhD Thesis, Polytechnic of Wales.

Schreinemakers, JF and Touretzky, DS, 1990, *Interfacing a Neural Network with a Rule-Based Reasoner for Diagnosing Mastitis*, proceedings of the International Joint Conference on Neural Networks, pp 487-90.

Wiltshaw, DG, 1991, Valuation by Comparable Sales and Linear Algebra, *Journal of Property Research*, Vol. 8, pp 3-19.

Wolverton, M and Diaz, J, 1996, *Investigation into Price Knowledge Induced Comparable Selection Bias*, RICS Cutting Edge.

Worzala, EM, Lenk, MM and Kinnard, WN, 1996, *The Impact of "Client Pressure" on the Appraisal of Residential Properties*, Paper Presented to Academy of Financial Services, New Orleans, LA, October.

A6.6 Almond, N.I., Lewis, O.M., Jenkins, D.H., Gronow, S.A. and Ware, J.A.

"Identification of Residential Property Sub-Markets Using Evolutionary and Neural Computing Techniques", Submitted to the Journal of Neural Computing and Applications, Jan. 1999.

Identification of Residential Property Sub-Markets Using Evolutionary and Neural Computing Techniques

Lewis O M, Ware J A & Jenkins D H.
School of Accounting and Mathematics, University of Glamorgan, Trefforest, Mid Glamorgan.

Abstract

This paper expands on previous work considering methods of stratifying property data in order to enhance its susceptibility to modelling for mortgage value estimation. Previous work (Lewis, et al, 1997) considered a clustering approach using a Kohonen Self-Organising Map (SOM) to stratify the training data prior to training a suite of MLPs. Although useful, this approach suffers from its estimation of trainability post clustering. The following method ameliorates the approach by replacing the static clustering step with a dynamic genetic algorithm implementation. The results show a healthy improvement in accuracy over the non-stratified approach and a more consistent level of accuracy compared with the Kohonen SOM approach. The paper concludes by analysing the underlying content of the derived stratas, thus providing a 'human readable' element to the approach that enhances its potential for acceptance by valuation institutions.

Introduction

The value of a residential property can be expressed in terms of property attributes (p), locational attributes (l) and other attributes (o) and hence value (v) can be estimated as v = f(l, p, o). Deriving this valuation function (f) has been the focus of a considerable amount of research, employing many different paradigms (Adair, et al, 1995; Wiltshaw; Evans, et al, 1992). One of the main problems concerns the effect of location in the appraisal process. Property has a fixed location and, therefore, its value will depend on the benefits of owning a property at that specific location. These benefits depend upon general environmental factors and specific local factors, the relationship between employment opportunities, communications and the general facilities of an area (Mackmin, 1994). Furthermore, the location of a property can effect the relative influence other attributes have on value. For example, proximity to a noisy airport effects the value of detached houses more than terraced houses (Evans, et al). To enable accurate modelling, researchers often select training properties from an homogeneous area allowing them to ignore location. However, this requires considerable a priori knowledge, a more effective approach is to include a stratification process prior to training (Adair, et al, 1995; Almy, 1997).

Heterogeneous Property Market

Before considering approaches for automating the stratification process, it is useful to develop a visual interpretation of the role of sub-markets in a heterogeneous market. Figure 3 provides a purely abstract view of the interplay of sub-market functions in a heterogeneous market. Here a heterogeneous market is viewed as a conceptual mathematical space containing many functions, accounting for the observed and described sub-market behaviour. The theoretical aim of any stratification process is then to segment this multi-functional space into smaller sub-regions containing a single value function.

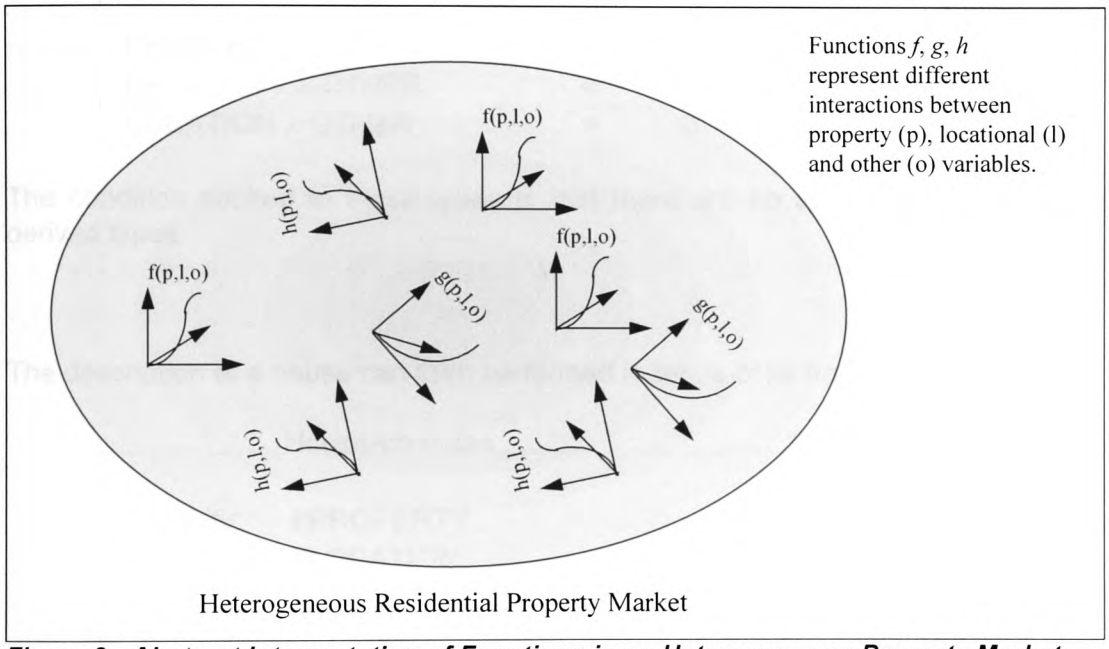


Figure 3 - Abstract Interpretation of Functions in an Heterogeneous Property Market Described in a Mathematical Conceptual Space.

Redefining the Problem Using Formal Methods.

To establish a structured approach to modelling an heterogeneous area, it is useful to formulate the problem in a precise way. This section models an heterogeneous property market using the formal specification method Z (See Spivey, 1989).

Given Types

In order to define an heterogeneous area in terms of homogeneous sub-regions and the houses within those sub-regions, it is useful to define 2 given types:

ATTRIBUTE: a set of attributes that can be used to describe a house;

PLACE: geographical identifier.

This is defined as:

[ATTRIBUTE, PLACE]

Attributes of a House

A house can be described in terms of a finite set of attributes. These attributes can be grouped into three main types: property (number of bedrooms, type, garage etc.); location (employment, schools etc.); and other (market trends etc.). This can be defined as:

$$\begin{array}{|l} \text{PROPERTY, LOCATION, OTHER: } \mathbb{P}\text{ATTRIBUTE} \\ \hline \text{PROPERTY} \wedge \text{LOCATION} = \varnothing \\ \text{PROPERTY} \wedge \text{OTHER} = \varnothing \\ \text{LOCATION} \wedge \text{OTHER} = \varnothing \end{array}$$

The condition applied to these types is that there are no attributes shared by the derived types.

The description of a house can then be formed in terms of its housing attributes:

$$\begin{array}{|l} \text{HouseAttributes} \\ \hline \text{property: } \mathbb{P}\text{PROPERTY} \\ \text{location: } \mathbb{P}\text{LOCATION} \\ \text{other: } \mathbb{P}\text{OTHER} \\ \hline \end{array}$$

Obviously, a house possesses more than just its description - it also has a value. A basic valuation function can be defined that maps housing attributes to value:

$$\text{BasicValuationFunction: } \text{HouseAttributes} \rightarrow \mathbb{N}$$

Theoretically, if all aspects of a house and their relative impact on value are known, this mapping should be true for all houses. However in practice, the derivation of such a mapping is currently beyond the scope of the available modelling tools. Hence, a more realistic approach is to define a valuation function for each homogeneous sub-region, and group together those sub-regions that share the same

valuation function. A sub-region can be defined in terms of its house types and its geographical location:

SUBREGION

theHouseTypes:	ℙHouseAttributes
thePlace:	PLACE

This allows the valuation function to be redefined as a higher order function, which when given a region will apply a function to the housing attributes and return the property value. This new valuation function is defined as:

ValuationFunction: SUBREGION → (HouseAttributes → ℕ)

This allows a particular house to be defined in terms of its attributes, region and value. Furthermore, the particular valuation function associated with the region when applied to the housing attributes returns the property value. This is defined as:

House

itsHouseAttributes:	HouseAttributes
itsValue:	ℕ
itsRegion:	SUBREGION

ValuationFunction (itsRegion) ((itsHouseAttributes) = itsValue)

Defining Sub-Regions by Stratifying the Heterogeneous Space

The aim of stratification is to identify the sub-regions that share the same valuation function and group these into a single model. The definition of a sub-region therefore depends on the factors that define the homegeneity of an area of which the most fundamental are location (ref) and property type (ref).

Stratification by Property Attributes

Here a sub-region is defined as a collection of houses sharing the same property attributes.

HomogeneousStrataByPropertyType

theSubRegions: ℙSUBREGION

$\forall S_1, S_2$: theSubRegions | S_1.thePlace $\neq S_2$.thePlace $\wedge$
S_1.theHouseTypes.Property = S_2.theHouseTypes.Property
$\forall h_1$: S_1.theHouseTypes, h_2: S_2.theHouseTypes •
ValuationFunction (S_1)(h_1) = ValuationFunction (S_2)(h_2)

Stratification by Locational Attributes

Here a sub-region is defined as a collection of houses sharing the same locational attributes.

HomogeneousStrataByLocation

theSubRegions: ℙSUBREGION

$\forall S_1, S_2$: theSubRegions | S_1.thePlace $\neq S_2$.thePlace $\wedge$
S_1.theHouseTypes.Location = S_2.theHouseTypes.Location
$\forall h_1$: S_1.theHouseTypes, h_2: S_2.theHouseTypes •
ValuationFunction (S_1)(h_1) = ValuationFunction (S_2)(h_2)

Practical Interpretation.

Unfortunately, due to the subjective and unique nature of property value, perfect stratification is not obtainable. However, sub-optimal stratification will undoubtedly go some way to improving the modelling capabilities. Two sub-optimal methods are considered in this paper. The first, which is decribed more fully elsewhere (Lewis & Ware, 1997) uses a Kohonen Self-Organising Map to cluster similar characteristics, followed by an investigation of the usefulnes of these clusters as training sets for an MLP (Section 4). The second approach uses Genetic Algorithms to generate random stratas which are refined in successive generations (Section 6).

Kohonen Self-Organising Map Approach

Geodemographic indicators could be employed to describe sub-regions in an heterogeneous residential area. Those sub-regions, sharing the same characteristics, could be grouped into a single model - based on the assumption that similar areas have similar underlying value functions. Extending this reasoning,

clusters found in Census data may correlate with homogeneous regions with respect to location, and clusters found in property data may describe homogeneous regions of properties. Early work by James (1994) concluded that an unsupervised neural network might be able to discern groupings within a parent data-set that might represent homogeneous areas.

Each observed cluster generated by the Kohonen SOM is used to train a single MLP network as shown in Figure 1.

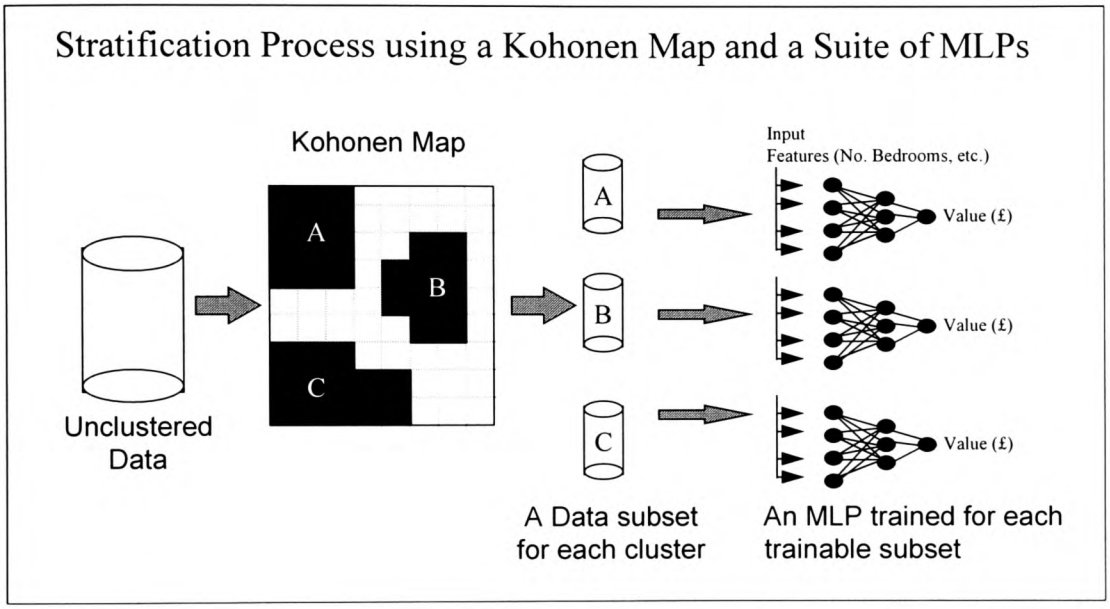


Figure 1 - Kohonen Stratification Approach.

The appeal of this approach is in its ability to estimate the trainability of each cluster using the Gamma test (Stefánsson, et al, 1997; see Lewis, et al, 1997 for a complete description of the process). This permits only 'useful' clusters to form training sets for MLP networks.

In order to provide a bench mark for analysing the methodology, a single MLP network was also trained on the whole data set. After training, the ability of the network to appraise residential properties with known values was tested, the results are shown in Table 1.

Table 1 - Results Obtained from Clustering Property Types

	Conventional Method	*Kohonen Method*
Mean Absolute % Error	*18%*	*8%*
% of Records with error > 10%	*74%*	*22%*
Minimum mean abs % error	*0%*	*0%*
Maximum mean abs % error	*310%*	*49%*

From the results obtained, it is evident that the methodology compares very favourably with the more conventional neural network approach, with the predictions made using the new method, 10% closer to the values returned by the valuer. This technique was also used to find clusters in Census data. This was achieved by using the Census variables as the training data for a Kohonen SOM. Results show an improvement in accuracy of between 1% and 14% using the sub-models over the single-model benchmark.

Acting against the method is its estimation of trainability post clustering. If care is not taken when choosing variables to use as inputs to the Kohonen map, then a high proportion of the clusters will not be suitable as training sets. For example, if a cluster relates to the variable FrontDoorColour = White then this will of course be unlikely to describe homogeneous properties. Although an absurd example, the reliance on a priori knowledge and post clustering fitness testing must be noted.

To overcome this problem, the Authors have investigated a further technique that permits fitness testing during clustering. This method recodes the market segmentation problem as a combinatorial one, before using Genetic Algorithms to produce homogeneous sub-sets.

Including a Fitness Metric in the Stratification Process.

The importance of a fitness metric is crucial if vector space W is to be split optimally providing the homogeneity required. Chen, et al (1997) provides a framework for stratifying a data set into trainable data sub-sets (stratification is achieved using linear discriminants and fitness is estimated by training an MLP network on the current state and evaluating whether the MLP network is 'good' or 'bad'). The data used to train a 'bad' network is further analysed using the stratification technique. This idea of divide-and-conquer is meritous, especially for large scale problems where data show signs of multiple functions. However, the computational requirement is large for the tree structure suggested and, furthermore, the success of the MLP - as a state

evaluation function - is relative to the network parameters chosen and the quality of the test set. A better approach is to estimate the lowest mean-square error (MSE) present in the data regardless of idiosyncrasies associated with a particular modelling technique. This can be achieved using the Gamma test (Stefánsson, et al, 1997).

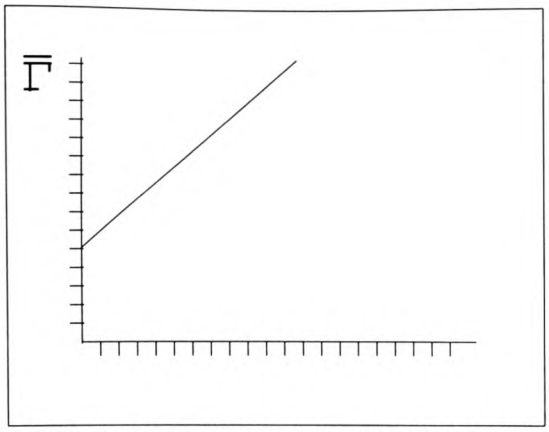
Gamma Test

The Gamma test attempts to estimate the best mean square error that can be achieved by any smooth modelling technique using the data. If y is the output of a function then the Gamma test estimates the variance of the part of y that cannot be accounted for by a smooth (differentiable) functional transformation. Thus if y = f(x) + r, where the function f is unknown and r is statistical noise, the Gamma test estimates Var(r).

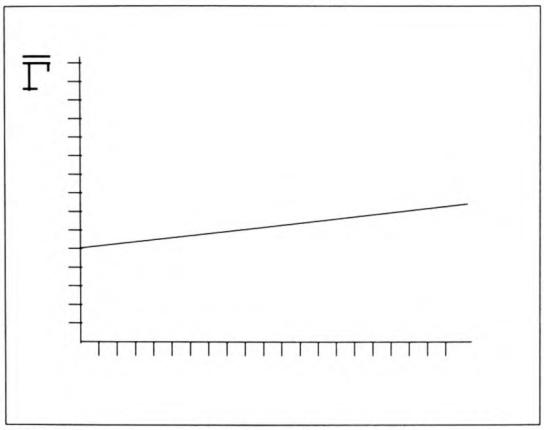
Var(r) provides a lower bound for the mean squared error of the output y, beyond which additional training is of no significant use. Therefore, knowing Var(r) for a data set allows prediction beforehand of what the MSE of the best possible neural network trained on that data would be [9].

Interpreting the output from the Gamma test requires considerable care and attention. The least squares regression line provides two pieces of information. First, the intercept on the Gamma axis is an estimate of the best MSE achievable by any smooth modelling technique. Second, the gradient gives an indication of the complexity of the underlying smooth function running through the data[8]. The Gamma test may estimate a very low MSE but unfortunately show a high level of complexity that could cause problems for a standard MLP network. It is easier to see this situation when the output from the Gamma test is presented graphically. A hypothetical example is shown in Figure 2 (a): with high noise content and high complexity; Figure 2(b): high noise and low complexity; Figure 2 (c): low noise and high complexity; and Figure 2 (d): low noise and low complexity (the desired outcome).

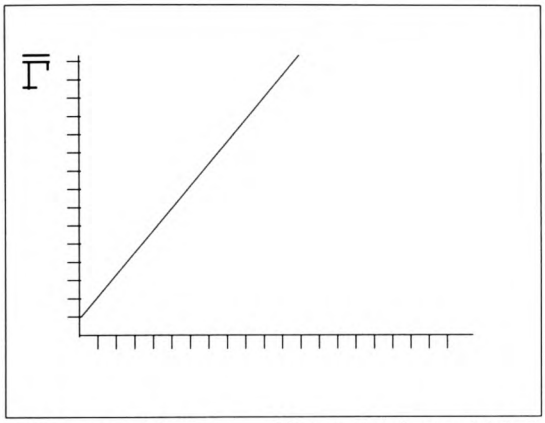
[8] This interpretation is based on emperical evidence and discussions with the research team who pioneered the Gamma test.



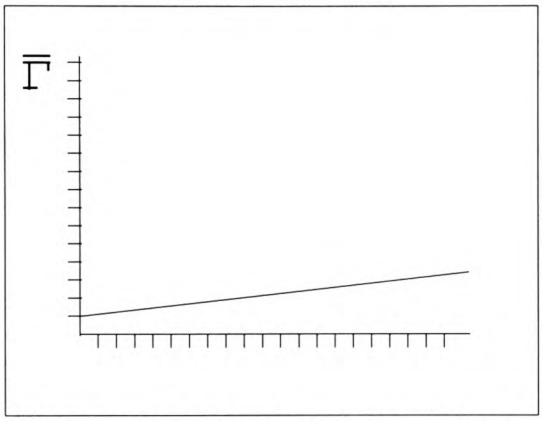
(a) High noise (large Γ value) and high complexity (steep gradient)



(b) High noise (large Γ value) and low complexity (flat gradient)



(a) Low noise (small Γ value) and high complexity (steep gradient)



(a) Low noise (large Γ value) and low complexity (steep gradient)

Figure 2

Utilising the Gamma Test

Chen's method (Chen, et al, 1997) could be adapted replacing the neural network state evaluation function (SEF) with the Gamma test. The estimated MSE of each state would then dictate whether further stratification was required. However, consider a scenario having 15 features subject to single binary partitions, the search space has an upper bound of 2^{15} (32768) possible states. The SEF will of course reduce the number of states but still the process is a large[9]. A popular alternative with such combinatorial problems is to recode the problem in a form suitable for Genetic Algorithm optimisation.

[9] It is the intention of the authors to investigate this approach for modelling residential property sub-markets and to compare the results gained with the results presented in this paper.

An Overview of Genetic Algorithms

Genetic Algorithms (GAs) simulate the Darwinian theory of evolution (See Goldberg, 1989 for a good introduction to GAs). A typical GA operates at the level of genetic coding: the chromosomes or genotypes. Genotypes are usually simple bit strings of fixed length; the related 'adult' individuals (phenotypes) are obtained by decoding such bit strings using domain information. The basic steps of a typical genetic algorithm are as follows:

1. Randomly generate a population of individuals (bit strings).
2. Decode each individual and evaluate its fitness.
3. Generate a new population using *cloning* (survival of current individuals); *cross-over* (bit string reproduction); and *mutation* (random changing of bits in current individuals).
4. Repeat steps 2 and 3 until convergence (or another stopping condition reached).

Mutation and Cross-over Operators

The cross-over and mutation operators are fundamental to the development of a GA solution, from its random initial state to a near optimal mature state. In most GA applications, the encoding permits the use of standard mutation operators (random inversion of bits in a chromosome) and standard single or multiple cut crossover operators. Figure 3 illustrates the effect of applying a mutation operator (a) and a single cut cross-over operator (b).

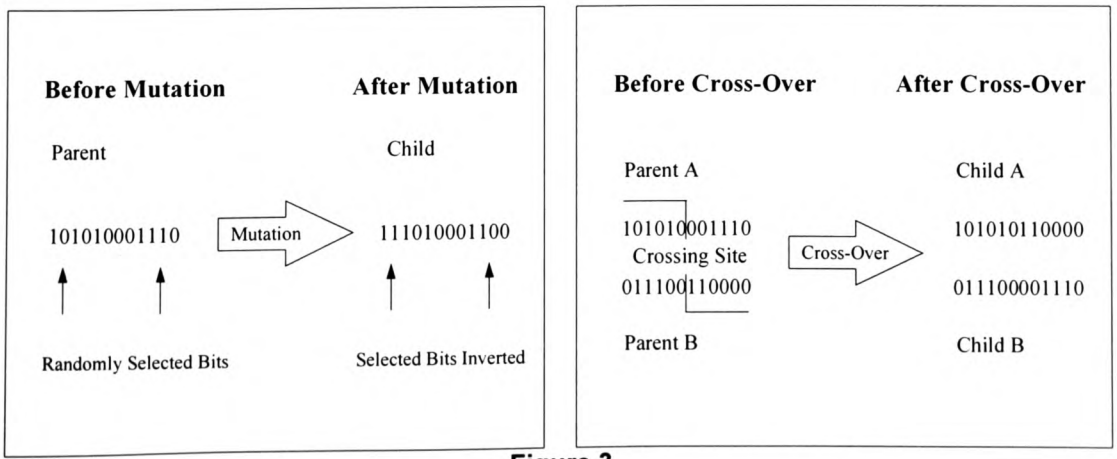


Figure 3

(a) A Schematic of simple mutation (b) A Schematic of simple cross-over

showing the inversion of randomly selected bits.

showing the partial exchange of information, using a crossing site chosen at random.

Stratifying Training Data Using a Genetic Algorithm

The GA approach benefits from having binary representations of the domain data. This allows the standard cross-over and mutation operators to be applied at bit level. The steps followed to formulate this stratification problem into one suitable for a GA to process were:

1. recode the Census data describing each ED as a bit string;
2. set up a look-up table to allow properties residing within an ED to be placed into a file suitable for analysis by the Gamma fitness test;
3. develop a GA application to generate stratas and test fitness.

Recoding the Census Data

The first step taken to achieve a binary representation for the 1991 Census data was to form a discrete representation of the raw data, by normalising between 0 and 100. For statistics relating to households, such as the number of terraced properties, this was achieved thus:

$$NormalisedValue = RawValue / NoOfHouseholds$$

For statistics relating to population, such as number of pensioners, the following equation was used:

$$NormailsedValue = RawValue / NoOfPersons$$

Each Census feature then represents percentage households or percentage persons in each ED. These discrete representations can be converted to binary representations by setting thresholds and using values of True (1) if the subject value is below the threshold and False (0) otherwise. For example, consider a Census variable C_1 describing the percentage of terraced properties in an ED, with thresholds of C_1<10% (low); 10%<= C_1< 20% (mid); C_1>=20% (high). An ED with 30% terraced properties could be described using 3 binary variables as: 0 (low) 0 (mid) 1 (high). These partitions can be improved by the addition of a fuzzy boundary, making the transition from one classification to the next less strict. This type of continuous

valued thresholding is sometimes known as soft partitioning. Figure 4 (a) and (b) give examples of a Census variable split by 2 and 3 soft partitions respectively.

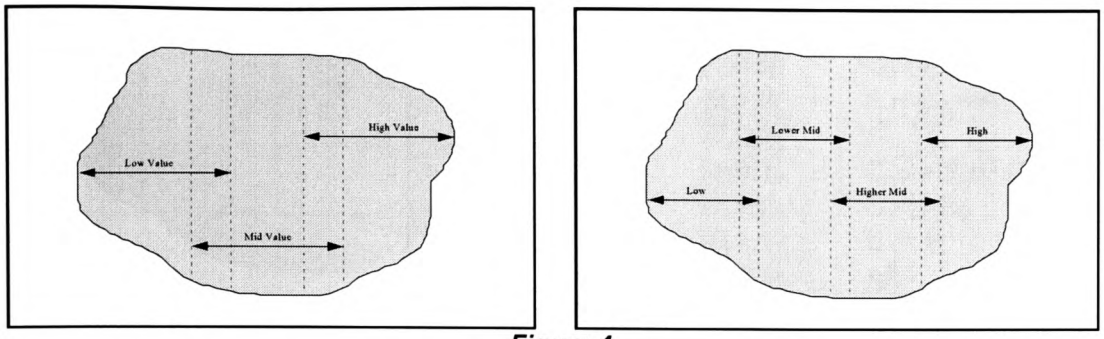


Figure 4

(a) 2 Soft Partitions **(b) 3 Soft Partitions**

Splitting each Census feature in this way allows a binary value to be used for each partition. For example in the case of the 2 partitions an ED that has a value below the 'Low' threshold can be represented as [1,0,0] and mid and high value EDs can be represented as [0,1,0] and [0,0,1] respectively.

The thresholds have been set manually after an inspection of the data. However, given the size of the task, an automatic approach was favoured. The following equations were used to set the thresholds for all Census variables given any number of partitions:

$$\text{LowerBound}_i = (i-1)*(\beta-\alpha) + Min \quad (1)$$

$$\text{UpperBound}_i = LowerBound_i + Width \quad (2)$$

Where: i: Partition Number

$Width = 2\alpha + \beta$

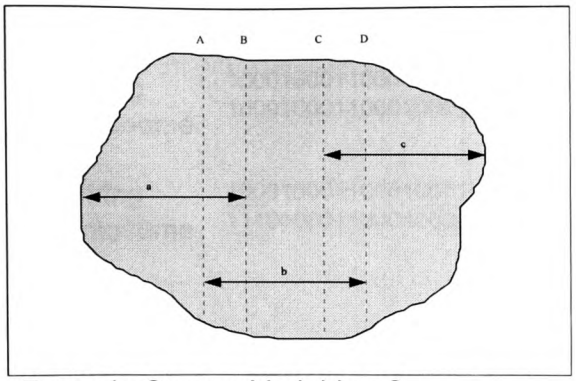
$\beta = \dfrac{Max - Min}{n}$

$\alpha = Overlap\ Ratio * \beta$

Max: Maximum Value for Feature

Min: Minimum Value for Feature

Figure 5 shows an example, where a Census variable has been divided using 2 partitions. All possible binary representations and their decoded meaning are given.



Example Census Variable : C_1

[a,b,c]	EDs Selected
[0,0,0]	All
[1,0,0]	C_1 <= B
[0,1,0]	A < C_1 <=D
[0,0,1]	C_1 > C
[1,0,1]	C_1 <= B or C_1 > C
[1,1,0]	C_1 <= D
[0,1,1]	C_1 > A
[1,1,1]	All

Figure 5 - Decoding the Binary Representations Found in an Example Using 2 Partitions.

Decoding and State Evaluation

Chromosome decoding is required after each new generation is created, in order to fix probabilities for survival (compete or partial) in later life. Decoding is relatively simple in this application with each (n-1) bits (where n is the number of partitions) representing a single Census variable. EDs that satisfy a chromosome description can be identified, and residential properties within the selected EDs can be grouped and placed into a separate data set. State Evaluation involves applying the Gamma test to the described data set and the appropriate metric(s) recorded. However, to speed up this process an initial pass of the Census data set is performed before the GA is run. The Census variables associated with each ED are converted into a binary representation using the described technique. Individual EDs can then be selected during the GA run by matching binary strings ignoring any [1,1,1] and [0,0,0] sub-strings in the current generation. Figure 6 provides examples of this partial matching technique between chromosomes and EDs.

	A B C D E F G 12312312312312312 3	7 Census variables (A..F) divided into 3 groups using 2 soft-partitions
ED String Chromosome	10001000110001000 1100 10001000110001000 1100	Chromosome matches ED string exactly so ED is selected
ED String Chromosome	10001000110001000 1100 11101000110001000 0100	Chromosome matches ED for each Census variable except where chromosome string is [1,1,1] or [0,0,0]. Hence, ED is selected
ED String Chromosome	10001000110001000 1100 10001000110101000 0100	Chromosome does not match ED string so ED is not selected.

Figure 6 - Illustration of Partial Chromosome/ED Matching Used to Select EDs on the Basis the Current GA Solution

Testing the Methodology

To test the methodology, residential property data from an heterogenous area was selected. Using the postcodes for each property, the enumeration district data was extracted from the national database.

Residential Property Data

A database containing information on residential property transactions during the period January 1993 to December 1995 was selected to test the methodology. There are 990 records in the original database, each with 51 attributes. However, a number of these have either constant values or free form text which is difficult to recode and were, therefore, removed. A description of the database used for this study is shown in Table 1.

Table 13 - A description of the database.

Attribute Name	Example Value	Attribute Name	Example Value
Street Name	Newport Road	Main Heating	Full, Partial, None
District or Village	Roath	Number of Bedrooms	1 - 8
Unit Type	Mid terraced etc.	Age in Years	0 - 500
Unit Size	Area M^2	Number of Garages	0 - 2
Valuation Date	19 May 1995	Value	10,000-255,000

1991 UK Census Data

Census data is available at a number of abstraction levels with the smallest being Enumeration District (ED) in England and Wales, and Output Area (OA) in Scotland. Originally EDs were intended to contain between 15 and 200 inhabited houses, with current ED sizes representing a workload that can be performed by a single enumerator in the time available, given the circumstances of the area. In the 1991 Census there were 106,866 ED's in England and 6,330 in Wales. A coding system allows access keys to be generated by following a country/county/district/ward/ED route. In addition to this a cross-reference from Postcode to ED allows individual properties to be included in ED level statistics (Dale and Marsh, 1993).

Ideally, to achieve maximum benefit from the Census data, the Cardiff dataset should be expanded at Postcode/ED level. As this is labour intensive, it is important that only useful information is extracted from the Census database. The empirical results described in this section used the Census data described in Table 3 (selected on the advice of valuation experts and literature searches).

Table 3 - Census Variables Used in Analysis

Socio-Economic Group	
Employers and Managers (Large est.)	Employers and Managers (small est.)
Professional workers (self-employed)	Professional workers (employees)
Ancillary workers and Artists	Foreman and Supervisors (non-manual)
Junior non-manual workers	Personal Services workers
Foreman and Supervisors (manual)	Skilled Manual workers
Semi-Skilled Manual workers	Unskilled Manual workers
Members of Armed Forces	
Employment	
Full-Time Employment	On Government Scheme
Part-Time Employment	Unemployed
Self Employed	
Qualifications	
Qualified Persons	Higher Degree
Degree	Diploma
Qualified and on Government Scheme	Qualified and Unemployed
Age Ranges of Qualified Persons	
Housing Stock	
Detached Properties	Purpose-Built Flats
Semi-Detached Properties	Converted Flats
Terraced Properties	Bedsits
Tenure	
Owner Occupied (Outright)	Owner Occupied (Buying)
Privately Rented (Furnished)	Privately Rented (Unfurnished)
Rented from Housing Association	Rented from Local Authority
Amenities	
Shared Use of WC	Exclusive Use of WC
Central Heating	
Availability of a Car	
Households with no car	Households with 1 car
Households with 2 cars	Households with 3+ cars
Ethnicity	
White	Black Caribbean
Black African	Black Other
Indian	Pakistani
Bangladeshi	Chinese
Asian	Persons born in Ireland
Miscellaneous Variables	
Working Mothers (Part-Time)	Working Mothers (Full-Time)
Lifestages (age ranges of residents)	Overcrowding (persons per household)
Travel to work estimates	Migration

Results

Table 4 shows the results obtained using the described methodology for each variable grouping. The mean absolute percentage error is shown for models trained on 'sub-market' data and also for the same data using a single MLP model. Any improvement in modelling accuracy is also shown.

Table 4 - Results Obtained for Census Variable Groupings Using the Described Methodology.

Census Variable Type	Gamma Interecept	Gamma Gradient	Single Model	Sub-Model	Improvement
Whole Dataset	0.0870	0.0284	20.57%	-	-
Residents Age	0.0289	0.0641	24.78%	14.05%	10.73%
	0.0302	0.0725	22.66%	16.69%	5.97%
	0.0307	0.0550	23.88%	14.89%	8.99%
	0.0196	0.1308	19.89%	17.73%	2.16%
	0.0227	0.1232	19.78%	18.3%	1.48%
Economic Position	0.0358	0.0929	20.53%	20.24%	0.29%
	0.0362	0.0953	20.65%	20.25%	0.45%
	0.0386	0.0915	20.60%	20.84%	-0.24%
	0.0398	0.0929	20.77%	21.28%	-0.51%
	0.0411	0.0946	20.80%	20.07%	0.73%
Amenities	0.0179	0.0845	19.23%	12.68%	6.55%
	0.0286	0.1085	21.94%	21.70%	0.24%
	0.0358	0.0751	21.68%	21.95%	-0.27%
	0.0452	0.0717	22.23%	23.10%	-0.87%
Car Availability	0.0043	0.1410	19.75%	15.52%	4.23%
	0.0224	0.0769	19.42%	15.65%	3.77%
	0.0268	0.0780	19.95%	14.32%	5.63%
	0.0253	0.0880	18.11%	12.20%	5.91%
	0.0216	0.1128	22.41%	16.24%	6.17%
Tenure	0.0005	0.1685	22.94%	20.35%	2.59%
	0.0156	0.1157	17.63%	14.72%	2.91%
	0.0119	0.1309	25.16%	20.35%	4.81%
	0.0112	0.1337	25.19%	20.00%	5.19%
	0.0218	0.0897	25.07%	19.62%	5.45%
Working Parents	0.0433	0.0490	19.82%	13.01%	6.81%
	0.0434	0.0489	19.81%	12.56%	7.25%
	0.0504	0.0889	20.77%	21.13%	-0.36%
Profession	0.0144	0.1024	20.54%	16.36%	4.18%
	0.0162	0.1014	19.78%	15.62%	4.16%
	0.0185	0.0965	20.52%	16.65%	3.87%
	0.0216	0.0941	27.60%	12.26%	15.34%
	0.0222	0.0946	26.49%	10.23%	16.26%
House Type	0.0077	0.1117	21.45%	14.64%	6.81%
	0.0135	0.1041	21.58%	20.55%	1.03%
	0.0172	0.0919	21.06%	14.80%	6.26%
	0.0061	0.1437	22.79%	8.29%	14.5%
	0.0183	0.1190	21.49%	17.15%	4.34%
Ethnic	0.0177	0.1317	21.45%	21.64%	0.19%
	0.0435	0.1041	21.58%	20.95%	-0.63%
Migration	0.0223	0.1111	19.35%	19.99%	0.64%
	0.0412	0.0945	21.03%	21.12%	0.09%
Travel to Work	0.0345	0.1223	23.04%	23.56%	0.52%

Discussion of Single Category Results

To determine whether the sub-model approach has been successful for the single category analysis, it is useful to revisit the results. Table 5 presents the average sub-model error compared with the error observed for the single control model.

Table 5 - Summary of Results for Single Census Category Sub-Models

Census Category	Single Model	Sub Model	Improvement
Residents Age	22.2%	16.3%	5.9%
Economic Position	20.7%	20.5%	0.2%
Amenities	21.27%	19.9%	1.37%
Car Availability	19.9%	14.8%	5.1%
Tenure	23.2%	19.0%	4.2%
Working Parents	20.1%	15.6%	4.5%
Profession	22.9%	14.2%	8.7%
House Type	21.7%	15.1%	6.6%
Ethnic	21.5%	21.3%	0.2%
Migration	20.2%	20.6%	-0.4%
Travel to Work	23.04%	23.56%	-0.52%

Clearly, some of the sub-models outperformed the single model by a significant margin. The results indicate an overall improvement in modelling accuracy for the sub-model approach compared to the single-model approach. The largest individual improvements using sub-models were observed for: Profession; House Type; Car Availability; and Residents Age. Significant improvements were also made in Tenure and Working Parent sub-models.

Although this type of selective analysis is of some benefit in ascertaining underlying model parameters, it is more probable that a neighbourhood description encompasses more than just one geodemographic feature grouping. Table 6 gives the results obtained when a selection of Census variables from different groupings were combined and used to generate sub-models.

Table 6 - Results Obtained Using a Selection of Census Variables.

Description	Gamma Intercept	Gamma Gradient	Single Model	Sub-Model	Improvement
A selection of Census data. (see Section X)	0.0131	0.1156	28.42%	14.57%	13.85%
	0.0116	0.1226	18.05%	15.57%	2.48%
	0.0206	0.0856	30.14%	16.05%	14.09%
	0.0208	0.0860	29.25%	15.90%	13.35%
	0.0105	0.1357	20.17%	18.92%	1.25%
	0.0114	0.1208	19.48%	18.78%	0.7%

Discussion of Results

Overall, the predictions made by the Census analysis sub-models outperformed those of single control model. An average increase in mean absolute predictive accuracy of 7.7% was observed.

To appreciate the composition of the sub-models, it is useful to examine the individual chromosomes that define each sub-model and to decode them back into Census aggregates:

Sub-Model 1

Percentage of population unemployed is in the range 0 - 2%

Ratio of rooms per person is in the range 2 - 4 rooms per person

Percentage of mortgaged properties is in the range 0 - 6%

Percentage of local authority rented properties is in the range 0 - 9%

Percentage of semi detached properties is in the range 0 - 7%

Comparing the ranges contained in this sub-model with the whole data set, this sub-model describes an area with: low unemployment; high rooms per person ratio; low number of mortgaged properties; low number of local authority rented properties; and, low number of semi-detached properties.

Sub-Model 2

Ratio of rooms per person is in the range 2 - 4 rooms per person

Percentage of mortgaged properties is in the range 5 - 12%

Percentage of semi-detached properties is in the range 0 - 7%

Percentage of terraced properties is in the range 6 - 13%

This sub-model describes an area with: high rooms per person ratio; an average number of mortgaged properties; low number of semi-detached properties; and, an average number of terraced properties.

Sub-Model 3

Ratio of cars per person is in the range 0.3 - 0.6 cars per person
Ratio of rooms per person is in the range 2 - 4 rooms per person
Percentage of mortgaged properties is in the range 0 - 6%
Percentage of local authority rented properties is in the range 0 - 9%

This sub-model describes an area with: an average ratio of cars per person; a high ratio of rooms per person; a low number of mortgaged properties; and, a low number of properties rented from a local authority.

Sub-Model 4

Ratio of cars per person is in the range 0.3 - 0.6 cars per person
Percentage of mortgaged properties is in the range 0 - 6%
Percentage of local authority rented properties is in the range 0 - 9%

This sub-model describes an area with: an average ratio of cars per person; a low number of mortgaged properties; and, a low number of properties rented from a local authority.

Sub-Model 5

Ratio of cars per person is in the range 0.3 - 0.6%
Percentage of local authority rented properties is in the range 0 - 9%
Percentage of terraced properties is in the range 0 - 6%

This sub-model describes an area with: an average ratio of cars per person; a low number of properties rented from a local authority; and, a low number of terraced properties.

Sub-Model 6

Percentage population unemployed is in the range 0 - 2%

Ratio of rooms per person is in the range 2 - 4 rooms per person

Percentage of local authority rented properties is in the range 0 - 9%

Percentage of semi detached properties is in the range 6 - 13%

This sub-model describes an area with: a low number of unemployed people; a high ratio of rooms per person; a low number of properties rented from a local authority; and, an average number of semi-detached properties.

Summary and Conclusions

At the outset of this research, two things were required:

- a means of estimating the modelling capabilities of a data-set, and;
- a method of sub-dividing heterogeneous difficult-to-model data into homogeneous sub-sets for which modelling performance is proven.

The technique employed to estimate the susceptibility of the data to be modelled was the Gamma test. Based on a nearest-neighbour approach, this method gives a measure of both noise (intercept) and complexity (gradient), assuming a single smooth continuous function underpins the data-set. A number of assumptions can therefore be made about a set of data given its Gamma results. Firstly, a data-set with a high noise value may have insufficient examples, descriptive features or contain data mapped by multiple underlying functions. Assumptions can also be made based on the complexity value, a very complex underlying function may in fact be an aggregate of multiple functions which may be too complex for MLP or MRA to model.

Two methods were investigated to achieve market segmentation, each using the Gamma test to measure the homogeneity of the generated sub-sets. First, a Kohonen SOM was used to cluster Census data and secondly, a Genetic Algorithm approach was investigated.

Kohonen SOM Market Segmentation (Lewis, et al, 1997)

Census aggregates were passed as input vectors to a Kohonen SOM that clusters data according to their cross-characteristic similarities. After training, each cluster represented a set of residential properties linked via an enumeration to postcode cross-reference look up table. The Gamma test was used to estimate the trainability of each subset, with those with low noise and low complexity forming training sets for individual MLP networks. The result show that the accuracy of the sub-models outperformed a single MLP control model within the range of 1 to 14%, with an average increase in modelling accuracy of 5%.

Although these result are promising, the generation of suitable training sets relies heavily upon selection of useful neighbourhood characteristics. Poor selection leads to clusters forming which perform badly when analysed by the Gamma test. This method is therefore of most use when a priori knowledge is available to determine the neighbourhood descriptors to select.

Genetic Algorithm Market Segmentation

To overcome this dependence on a priori feature selection, a second method was investigated that progresses to a solution iteratively by changing the features used to define a sub-model over successive generations. This method, implemented as a Genetic Algorithm randomly generates a number of sub-data-sets and tests their fitness using the Gamma algorithm. Those sub-sets (or at least very similar ones) that fair the best are more likely to appear in successive generations. This process is iterated for a number of generations with an elite population, formed from sub-data-sets with the best Gamma results, being the output from the process. After termination, the elite sub-sets are used to train individual MLP networks in the same manner as the Kohonen SOM approach.

Here, the results are more consistent than the Kohonen approach as fitness is tested on-route as opposed to post-clustering. An average increase in accuracy of 7.5% was observed, within a range of 0%- 16%. Some models made no aggregate improvement over the single control model with a few models fairing marginally worse.

References

Adair, A, McGreal, S, 1995, Investigation of the Influence of Property and Socio-Economic Variables on residential Values and the Formulation of Valuation Models Based on Regression Analysis, Technical Report, Real Estate Studies Unit, School of the Built Environment, University of Ulster (April 1995).

Almy, R, Horbas, J, Cusack, M, Gloudemans, R, 1998, The Valuation of Residential Property using Regression Analysis, Computer Assisted Mass Appraisal: An International Review, Ed. McCluskey, WJ and Adair, AS, Ashgate Publishing Company, England.

Chen, Ke, Xiang, Yu, Huisheng, Chi, 1997, Combining Linear Discriminant Functions with Neural Networks for Supervised Learning, Journal of Neural Computing and Applications, Vol. 6, Springer-Verlag, London.

Dale A, and Marsh, C, 1993, The 1991 Census User's Guide, HMSO Publications

Evans, A, James, H, Collins, A, 1992, Artificial Neural Networks: an Application to Residential Valuation in the UK, Journal of Property Valuation and Investment, Vol. 11, pp 195-204.

Goldberg, DE, 1989, Genetic Algorithms in Search Optimisation and Machine Learning, Addison Wesley

James, H, 1994, An 'Automatic Pilot' for Surveyors, RICS Cutting Edge.

Lewis, O.M., Ware, J.A., Jenkins, D.H, A Novel Neural Network Technique for Modelling Data Containing Multiple Functions, in Computational Intelligence - Theory and Applications, ed. Bernd Reusch, (Lecture Notes for Computer Science Series Vol. 1226), Springer Verlag, ISBN 3-540-62868-1, pp 141-149.

Mackmin, D, 1994, The Valuation and Sale of Residential Property, 2nd Edition, Routledge.

Spivey, J.M., 1992, The Z Notation: a Reference Manual, Prentice Hall International, Second Edition.

Stefansson, A, Koncar, N, Jones, AT, 1997, A Note on the Gamma Test, Journal of Neural Computing and Applications, Vol.5 No. 3, Springer Verlag

Waggert, S, 1997, Gamma Test Documentation available from Prof. A. Jones at the Department of Computer Science, University of Wales, Cardiff, UK.

Wiltshaw, DG, 1991, Valuation by Comparable Sales and Linear Algebra, Journal of Property Research, Vol. 8, pp 3-19.